المعهد الملكي للثقافة الأمازيغية
ⵜⴰⵙⴷⴰⵡⵉⵜ ⴰⵔⵉⵖⴰⵍ ⵏ ⵜⵣⵔⴼⵜ ⵜⴰⵎⴰⵣⵉⵖⵜ
INSTITUT ROYAL DE LA CULTURE AMAZIGHE

Le Centre des Etudes Informatiques, des Systèmes d'Information et de Communication

Actes de conférence

7ème Conférence Internationale

ⵜⴰⵎⴰⵣⵉⵖⵜ ⴷ ⵜⵜⵓⵔⵉⵡⵉⵜⵜ ⵏ ⵜⵣⵔⴼⵜ ⵏ ⵜⵙⵙⵓⵎⵉ ⴰ ⵜⵎⵣⵡⵓⵔⵓ

الأمازيغية وتكنولوجيا المعلومات والتواصل

Technologies d'Information et de Communication pour l'Amazighe

Coordination

BOULAKNADEL Siham

ⵜⴰⵎⴰⵣⵉⵖⵜ ⴷ ⵜⵜⴽⵏⵓⵍⵓⵊⵢⵜ ⵏ ⵓⵙⵎⵔⵙ ⴷ ⵓⵎⵙⵙⵓⵍⴻ

الأمـازيغية وتكنولوجيـا المعلـومـات والتـواصـل

Technologies d'Information et de Communication pour l'Amazighe

المعهد الملكي للثقافة الأمازيغية
ⵓⵔⵉⵔⵓ ⵓⵅⵏⴰⵡ ⵏ ⵜⵙⴷⵓⵍⵜ ⵜⴰⵎⴰⵣⵉⵖⵜ
INSTITUT ROYAL DE LA CULTURE AMAZIGHE

Centre des Etudes Informatiques,
des Systèmes d'Information et de Communication

Actes de conférence

7ème Conférence Internationale

ⵜⴰⵎⴰⵣⵉⵖⵜ ⴷ ⵜⵣⵔⵉⵡⵉⵏⵉⵣⵢⵜ ⵏ ⵙⵍⵖⵎⵣⵢⵜ ⴰ ⵙⵎⵙⵓⵍⵓⵖ

الأمازيغية وتكنولوجيا المعلومات والتواصل

Technologies d'Information et de Communication pour l'Amazighe

Coordination
Siham Boulaknadel

Publications de l'Institut Royal de la Culture Amazighe
Centre des Etudes Informatiques, des Systèmes d'Information et de Communication
Série : Colloques et séminaires N° 52

Titre
Technologies d'Information et de Communication pour l'Amazighe

Coordination
Siham Boulaknadel

Conception
Nadia Kiddi (Unité de l'édition)

Editeur
Institut Royal de la Culture Amazighe

Imprimerie
Imprimerie des Editions OKAD - Rabat - 2020

Dépôt légal
2015 MO 1163

ISBN
978-9954-28-184-0

ISSN
2421-9711

Copyright
©IRCAM

PREFACE

1. Introduction

Les technologies des langues sont une composante majeure de la société de l'information, qu'elles soient disponibles sur le marché ou en cours de développement, elles offrent des ressources inestimables pour tous ceux qui ont à produire, transformer ou rechercher des données. Cependant, les avancées en technologies des langues ne concernent qu'1% des langues parlées dans le monde, dont peu de langues minoritaires, en particulier celles du Maroc. Pourtant, quelques expériences (langues basque et catalane en Espagne entre autres) nous révèlent que réunir volonté politique, savoir scientifique et savoir-faire technique permet le développement rapide des technologies des langues minoritaires en les dotant de ressources et d'outils nécessaires. Ceci permettra d'encourager des recherches plus fondamentales sur les langues minoritaires et de s'attaquer au défi de l'élaboration d'un multilinguisme véritable considérant les variétés dialectales, tout en permettant la réalisation d'applications à forte valeur ajoutée pour les collectivités locales.

La conférence TICAM, qui s'est tenue à l'IRCAM du 28 au 29 novembre 2016, a pour objectif de faire tous les deux ans le point sur les résultats de recherches, les nouvelles applications, les retours d'expériences dans le domaine de l'informatisation des langues peu dotées. Cette septième édition de la conférence TICAM, qui a réuni des linguistes, des spécialistes du traitement automatique des langues et des industriels a permis de dresser le bilan des travaux de recherche entamés dans le domaine de la technologie d'information et de communication appliquée aux langues naturelles et particulièrement à la langue amazighe et prospecter les démarches à entreprendre dans l'avenir pour leur intégration dans la société de demain.

2. Thèmes

Plus que jamais, les technologies des langues est un atout décisif de l'insertion, la participation et le maintien des langues dans la société d'information. Les sciences et technologies de l'information et de la communication offrent de nouvelles et nombreuses opportunités pour la préservation et la revitalisation des langues ainsi que pour la pédagogie et l'apprentissage. Ces nouvelles technologies apportent aussi avec elles de nouveaux problèmes (conception et modélisation des apprenants et des connaissances, adaptation dynamique, etc.) et leurs évolutions demandent de constants réajustements (accès mobiles aux connaissances, plateformes massives d'enseignement sur le web, nouveaux outils de traitement numérique). La conférence TICAM aborde conjointement :

- les recherches scientifiques et innovations technologique en matière des technologies des langues ;
- l'apprentissage des langues induit par l'usage des nouvelles technologies ;
- les applications et les retours d'expériences.

TICAM 2016 a souhaité être l'occasion de présenter les projets et les travaux des scientifiques, des enseignants de toute discipline ainsi que des industriels proposant des solutions innovantes pour le traitement automatique des langues ainsi que pour leur apprentissage électronique.

3. Programme de l'édition 2016

Le programme de cette année présente trois conférenciers invités :

- Azzedine Mazroui (Université Mohamed 1er, Oujda) spécialiste en traitement automatique de l'arabe,
- Driss Marjane (Université Dhar El Mahraz, Fès) spécialiste en apprentissage médiatisé par la technologie,
- et Mohamed Ismail (Transsion Holding, Egypte) spécialiste en localisation des logiciels.

Le programme des contributions sont organisés en 4 grands thèmes :

- Ressources et outils linguistiques,
- Apprentissage et enseignement médiatisé par la technologie,
- Traitement de la parole,
- Reconnaissance optique des caractères.

En complément de ce programme, une session poster a été programmée pour donner l'opportunité aux jeunes chercheurs de présenter leurs travaux à la communauté.

4. Remerciements

Nos remerciements chaleureux vont tout d'abord aux auteurs pour leurs contributions scientifiques. Nous remercions aussi les membres du comité de lecture, pour le temps consacré, pour leur patience et leur bonne volonté.

Nos vifs remerciements vont également à toute l'équipe du comité d'organisation qui a mis tout en œuvre pour pouvoir permettre le meilleur accueil et le meilleur déroulement de la conférence. Enfin, nous remercions spécialement pour son soutien financier et aides diverses l'Institut Royal de la Culture Amazighe. Sans son soutien, ni la conférence TICAM 2016, ni ce recueil n'auraient vu le jour.

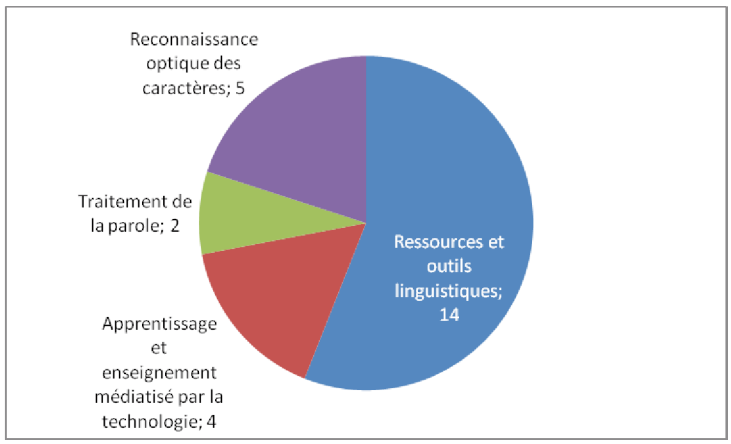
TICAM en chiffres ……

25 articles acceptés :

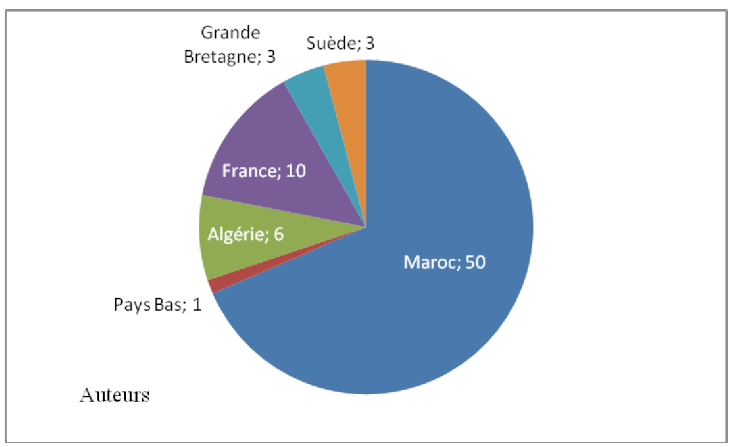
17 Présentations orales

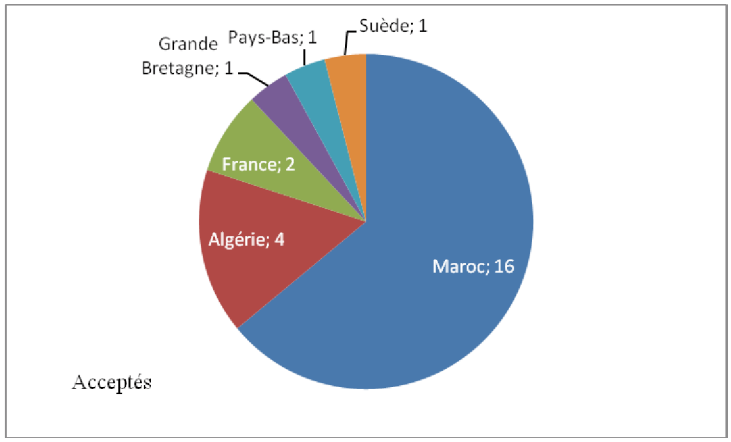
8 Posters

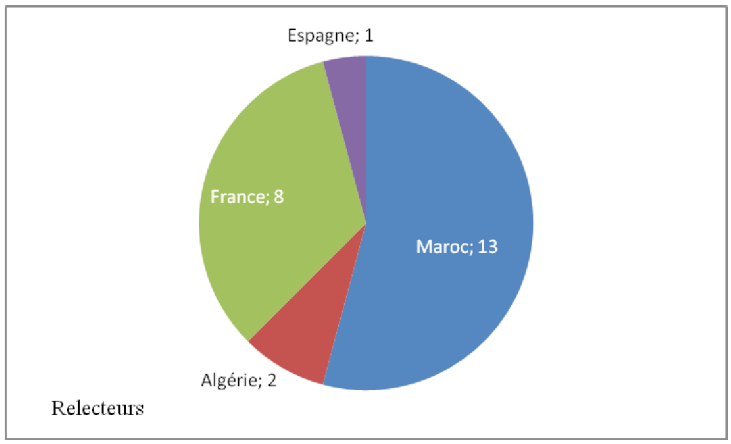
Programme consolidé



Répartition par pays




Grande Bretagne; 1
Pays-Bas; 1
Suède; 1
France; 2
Algérie; 4
Maroc; 16
Acceptés


Espagne; 1
France; 8
Maroc; 13
Algérie; 2
Relecteurs

Candidats au prix du meilleur article scientifique

Vers la construction des ressources linguistiques nécessaires pour la génération de la langue amazighe à partir de l'interlangue UNL
I.Taghbalout, F. Ataa Allah, M. El Marraki

Reconnaissance des caractères Tifinagh imprimés par une déscription structurelle
Y. Ouadid, B. Minaoui, M. Fakir

Vers un module d'analyse morphologique de la langue amazighe en utilisant la plateforme Nooj
F.Z. Nejme, S. Boulaknadel, D. Aboutajdine

Gagnant du prix du meilleur article scientifique

Reconnaissance des caractères Tifinagh imprimés par une description structurelle
Y. Ouadid, B. Minaoui, M. Fakir

Candidats au prix du meilleur poster

L'analyse morphologique de la langue amazighe
R. Ammari, L. Zenkouar, M.Outahajala

Etiquetage morphosyntaxique de la langue amazighe
S. Amri, L. Zenkouar, M.Outahajala

Corpus multilingues pour les langues peu dotées
N. Miftah, F. Ataa Allah, I. Taghbalout

Gagnant du prix du meilleur poster

Corpus multilingues pour les langues peu dotées
N. Miftah, F. Ataa Allah, I. Taghbalout

Comité scientifique

- Aboutajdine Driss (FSR, Rabat)
- Ansar Khalid (IRCAM, Rabat)
- Aoughlis Farida (UMMTO, Tizi Ouzou)
- Ait Kerroum Mounir (ENCG, Kénitra)
- Ataa Allah Fadoua (IRCAM, Rabat)
- Bennani Samir (EMI, Rabat)
- Bouikhalene Belaid (FST, Béni Mellal)
- Boulaknadel Siham (IRCAM, Rabat)
- Bouyakhf El Houssine (FSR, Rabat)
- Cavalli Sforza Violetta (AUI, Ifrane)
- Drissi Moulay M'hammed (MEN, Rabat)
- El Hannach Mohamed (USMBA, Fès)
- El Marraki Mohamed (FSR, Rabat)
- El Qadi Abderrahim (EST, Meknes)
- Fadouli Nouredine (EMI, Rabat)
- Fadili Hammou (CNAM, Paris)
- Fakir Mohamed (FST, Béni Mellal)
- Ichou Benaissa (IRCAM, Rabat)
- Idrissi Najlae (FST, Béni Mellal)
- Khalidi Idrissi Mohamed (EMI, Rabat)
- Mazroui Azzeddine (FS, Oujda)
- Mouradi Abdelhak (MESRSFC, Rabat)
- Morin Emmanuel (LINA, Nantes)
- Pognan Patrice (INALCO, Paris)
- Rami Salim (FPS, Safi)
- Rosso Paolo (UPV, Valence)
- Rguig Mohand (USMBA, Fès)
- Semmar Nasredine (CEA, Paris)
- Sidir Mohamed (UPJV, Amiens)
- Silberztein Max (UFC, Besançon)
- Souifi Hamid (IRCAM, Rabat)
- Tigziri Noura (UMMTO, Tizi Ouzou)
- Yousfi Abdellah (FSJES, Rabat)
- Zerouali Mohamed (CFIE, Rabat)
- Zock Michael (Univ-Marseille, Marseille)

Comité d'organisation

Centre des Etudes Informatiques des Systèmes d'Information et de Communication

Table des matières

Conférences invitées **15**

Apports et limites des outils du TALN face au Big data **15**
A.MAZROUI

Numérisation des langues africaines **16**
M.ISMAIL

Les technologies de l'information et de la communication et l'enseignement des langues: les défis et les opportunités **17**
D. MARJANE

Chapitre 1 : Ressources et outils linguistiques

Analyse syntaxique automatique en tamazight (Kabyle) **27**
F. YAMOUNI

Vers un module d'analyse morphologique de la langue amazighe en utilisant la plateforme Nooj **39**
F.Z. NEJME, S. BOULAKNADEL, D. ABOUTAJDINE

Etiquetage morphosyntaxique de la langue amazighe **49**
S. AMRI, L. ZENKOUAR, M.OUTAHAJALA

Vers la construction des ressources linguistiques nécessaires pour la génération de la langue amazighe à partir de l'inter-langue UNL **63**
I. TAGHBALOUT, F. ATAA ALLAH, M. EL MARRAKI

Corpus multilingues pour les langues peu dotées **77**
N. MIFTAH, F. ATAA ALLAH, I. TAGHBALOUT

Analyse textuelle de la ponctuation à l'aide du concordancier ANTONC d'une œuvre de Bélaid At Ali **89**
N.TIGZIRI

Convertisseur numérique : Tifinaghe – Braille **99**
N. YACOUBI, J. FRAIN, F. ATAA ALLAH

Arabicized and Romanized Berber Automatic Identification **109**
W. ADOUANE, N. SEMMAR, R.JOHANSSON

L'analyse morphologique de la langue amazighe..... 115
R. AMMARI, L. ZENKOUAR, M.OUTAHAJALA

Détection Automatique de fin des phrases pour la langue amazighe 127
M. BINIZ, B. El AYACHI, M. FAKIR

Le Machine Learning : numérique non supervisé et symbolique peu supervisé, une chance pour l'analyse sémantique automatique des langues peu dotées..... 137
H. FADILI

Etat de l'art de la traduction automatique des langues – Approches & Méthodes 149
N.BOUHRIM

IT Communication as institutional promotion for the technological realizations 159
M. BOUMEDIANE

La correction automatique d'orthographe dans la langue arabe..... 169
M. NEJJA, A. YOUSFI

نحو معجم حاسوبي للمتلازمات اللفظية في اللغة العربية المعاصرة 179
أ. الغامدي، إ. أتويل

Chapitre 2 : Apprentissage et enseignement médiatisé par la technologie

Tamazight E-Learning platform: Leveraging graph databases for flexible design..... 197
V. CAVALLI-SFORZA, A. ISSAR, M. OUAKRIME, A. AGNAOU, H. DARHMAOUI

Vers une démarche de qualité et de bonnes pratiques enseignantes des langues pour les filières scientifiques . Cas de l'intégration d'outils de simulation ludique et collaborative dédiés à l'apprentissage des langues naturelles 209
A. BOUKELIF

E-conte au service de l'apprentissage de l'amazighe..... 217
S. BOULAKNADEL, F. ATAA ALLAH

Perception évolutive envers les serious games et à la création vidéoludique par des enseignants stagiaires de la langue amazighe 225
I. OUAHBI, H. DARHMAOUI, F. KADDARI

Chapitre 3 : Traitement de la parole

Instrumentations pour l'étude des consonnes géminées du tarifit 233
F. BOUAROUROU, B. VAXELAIRE, Y. LAPRIE

Etablissement des corpus et construction de l'objet pour l'analyse de discours. Transcription synchronisée à l'aide du logiciel Winpitch 249
R.BOUKHERROUF, N. TIGZIRI

Chapitre 4 : Reconnaissance optique des caractères

Reconnaissance des caractères Tifinagh imprimés par une description structurelle 257
Y. OUADID, B. MINAOUI, M. FAKIR

Towards Internet domain names in Tifinagh 269
A. AIT ALI, A. SEDRATI

A Dataset of Amazigh printed words images 281
N.AHARRANE, A.DAHMOUNI, K. EL MOUTAOUAKIL

Toward Handwritten recognition using canonical representation for Multilingual Documents 291
Y. EL MALKI, Y. ES-SAADY, D. MAMMASS

Apports et limites des outils du TALN face au Big data

L'évolution de l'informatique a toujours été accompagnée de l'apparition de nouveaux défis. Ainsi, le développement du web et la large utilisation des réseaux sociaux ont permis l'émergence d'une connaissance collective ignorant les frontières et produisant en continu un nombre très grand de données.

Les contraintes relatives à la croissance exponentielle du contenu numérique sur le web (Big data) et les phénomènes d'ambiguïté qui l'accompagnent rendent son analyse par les outils du traitement automatique des langues naturelles (TALN) très délicate. Cela met les professionnels du domaine des technologies de l'information devant d'énormes défis.

Après avoir rappelé l'évolution du Bigdata et quelques-unes de ses applications, nous discuterons la problématique de l'ambiguïté des données textuelles et les différents niveaux de difficultés (morphologique, syntaxique et sémantique) qu'elles engendrent lors du traitement automatique de ces données. Ces ambiguïtés sont amplifiées dans le cas de la langue arabe à cause de l'absence des signes diacritiques dans l'immense majorité des textes (selon certaines études, plus de 98% des textes arabes ne sont pas voyellés). Nous présenterons les résultats d'une étude qui a montré que la principale source d'ambiguïté lors du traitement automatique des textes arabes est conséquence de l'absence de ces signes. De plus, cette étude a évalué le coût supplémentaire causé par l'absence des signes diacritiques lors de l'analyse automatique des textes arabes.

Nous terminons la conférence en exposant quelques techniques permettant de soulever certains cas d'ambigüité.

Azzedine Mazroui
PES, Faculté des Sciences d'Oujda

A propos de Azzedine Mazroui

Professeur de l'Enseignement Supérieur à la faculté des Sciences d'Oujda. Il est membre du Laboratoire de Recherche en Informatique de l'Université Mohammed Premier et directeur de l'équipe Traitement automatique des langues naturelles.

Ses recherches portent sur le traitement automatique de la langue arabe. Membre dans plusieurs projets internationaux.

Auteur de nombreux chapitres de livres et articles dans des revues de renom.

Numérisation des langues africaines

A propos de Mohamed ISMAIL

Directeur régional de développement des affaires pour le groupe Transsion. Plus de 19 ans d'expériences dans des sociétés multinationales spécialisées dans le domaine des télécommunications. Grande expérience sur le marché nord-africain.

M. Mohamed Ismail a occupé plusieurs postes dans la conception de réseaux, les services après-vente, la gestion de projet, les ventes et le développement des affaires.

Auparavant :

- Directeur des ventes OT compte chez Huawei Technologies
- Gestionnaire de comptes nord-africain chez Lucent Technologies

Le groupe TRANSSION, créée en juillet 2006, est une société de haute technologie spécialisée dans la recherche et le développement, la production, la vente et le service des produits de communication mobile. Au cours des dernières années, un intérêt particulier a été consacré à l'Afrique, notamment par la localisation des langues africaines, telles que le Swahili et l'Amharique. Aujourd'hui, le groupe compte dans son actif 31 langues localisées. Cet intérêt ne se limite pas aux produits de télécommunication, mais il s'étend aux ressources éducatives basées sur la haute technologie. Le groupe TRANSSION a réalisé des produits éducatifs modernes et innovants dotés des dernières technologies de pointe dans la perspective d'appuyer les programmes d'enseignement.

Mohamed ISMAIL,
Directeur régional de développement des affaires pour le groupe Transsion

Les technologies de l'information et de la communication et l'enseignement des langues: les défis et les opportunités

Résumé

Il fut un temps ou le bain linguistique familial était le seul foyer qui assurait la transmission et la préservation de la langue-mère. Dans le cas particulier du Tamazight cette fonction semble être une vocation naturelle des mères, à tel point que cette langue ait bien mérité son titre de *langue des mères* plutôt que *langue-mère*. De nos jours, et cela depuis les deux dernières décennies du siècle dernier, avec l'avènement de l'ère numérique et surtout celle du web social (web 2.0) et puis la naissance de ceux qu'on appela dès lors les natifs numériques, le capital culturel de cette nouvelle génération n'a pas cessé de connaître des bouleversements successifs majeurs. Du coup, les modes et les moyens de socialisation et d'acquisition des savoirs – et des langues - ont basculé vers ce que l'on appelle désormais la disposition technologique. Ces développements représentent des défis de grande taille pour le Tamazight qui reste une des langues les moins dotées au monde. Toutefois, ces mêmes développements offrent des opportunités inégalées, et autrefois inimaginables, pour une revitalisation et une plus large diffusion de cette langue. Il est à noter même que si l'intégration progressive du Tamazight aux nouvelles TIC n'a pas des effets immédiats sur les attributs sociolinguistiques de cette langue, Il n'en reste pas moins que cette intégration aura l'immense avantage de refaçonner le Tamazight au moule des temps modernes.

A propos de

Driss Marjane

Professeur de linguistique générale à l'Université Sidi Mohamed Ben Abdellah à Fès. Il est également administrateur de la plateforme d'enseignement à distance de la Faculté des Sciences Dhar El Mahraz, Fes. Driss Marjane est certifié Spécialiste en E-Learning par Qualilearning-Widdoo, Swisse et le Campus Virtuel Marocain en 2010. Depuis, Driss Marjane a effectué plusieurs formations en e-learning et en numérisation des contenus didactiques au Maroc et à l'étranger (Etats-Unis, 2011; Belgique, 2014 ; France, 2015). Driss Marjane a aussi publié des articles et il a animé des ateliers de formations et des communications sur les outils numériques au service de l'enseignement dans des rencontres nationales et internationales.

1. Introduction

Pour quelqu'un qui ne parle le Tamazight comme moi, j'avoue que je me sens un petit peu embarrassé d'aborder le sujet de la revitalisation de la langue Amazigh, tout en sachant bien que je ne suis ni entièrement Arabe, ni entièrement Amazigh. Mais, après tout, Max Weber (1968:389) avait bien définit les groupes ethniques comme ceux qui " […] entretiennent une croyance subjective en une descendance commune.» et il souligne « […] qu'il n'importait guère si oui ou non une relation objective de sang existait.»Et, bien que les communautés ethniques ne soient que le produit de notre imagination, elles ne sont pas totalement fausses, parce que se sont nos perceptions subjectives qui définissent les limites de notre identité collective. Et mes perceptions, en partie subjectives, me disent que je suis Marocain et Amazigh. Et loin de tous les discours revendicatifs qui se développent autour de la revitalisation du Tamazight, le sujet de ma communication est de parler des nouvelles technologies de l'information et de la communication et comment elles pourraient éventuellement contribuer au mieux à l'enseignement des langues, et dans le cas particulier du Tamazight, j'oserais dire comment elles pourraient contribuer à sa survie.

Dans cette communication, je vais dans un premier temps mettre face à face les modes traditionnels et les modes modernes, surtout numériques, de transmission de la langue et de la culture. Puis je passerai en revue un certain nombre de traits saillants du Web social qui semblent correspondre aux traits comportementaux de cette nouvelle génération que l'on dénomme « les natifs numériques » pour arriver, enfin, à la conclusion que toutes les langues qui se veulent modernes de nos jours doivent se fondre et couler dans ce moule qui est l'outil numérique pour en ressortir avec la vitalité de l'ère moderne.

2. Les modes et les mondes traditionnels d'acquisition des langues

Revenons un petit peu en arrière pour revisiter le contexte le plus naturel ou se déroulait la transmission de la langue. Il fut un temps – et il l'est encore dans bien des régions, surtout celles qui sont éloignées des grands centres urbains – ou le bain linguistique familial était le premier, sinon le seul foyer qui assurait la transmission et la préservation de la langue-mère. Dans le cas particulier du Tamazight, cette fonction était parfois une des taches réservées exclusivement aux mères à tel point que cette langue ait bien mérité son titre de langue des mères plutôt que langue-mère. L'éminente linguiste Marocaine Fatima Sadiqi décrit le statu du Tamazight, "en tant que langue d'identité culturelle, du foyer, de la famille, de l'appartenance au village, de l'intimité, des traditions, de l'oralité et de la nostalgie à un passé lointain. »Et elle continue en soulignant que « [Le Tamazight] perpétue les attributs qui sont considérés comme féminines dans la culture marocaine » (Sadiqi, 2003). Les commentaires de Sadiqi démontrent le rôle important que les femmes assument - ou dirais-je, assumaient - dans la continuité et dans la préservationnon seulement de la langue mais de l'ensemble du savoir culturel Amazigh. Dans une autreétude, documentaire cette fois(Synthia Becker, 2006), on peut lire ce qui suit : « Les mères passent les savoir-faire du tissage, de la broderie, du tatouage, de la poterie à leurs filles. Les femmes Amazighs démontrent l'estime, le respect, et le statut accordé aux mères en incorporant les symbolismes de la fertilité dans les tapis qu'elles tissent, dans leurs habits, leurs tatouages et leurs coiffures des cheveux. Par conséquent, les arts Amazighs sont des métaphores de maternité qui démontrent le rôle

crucial que les femmes jouent dans la transmission et la préservation de l'identité Amazigh. » Et pourtant, en dépit de ce rôle capital que la femme assume dans le processus de socialisation et de perpétuation de la langue et de la culture, il est à noter non sans amertume comme le note le linguiste Moha Ennajique « les langues-mères en générale sont marginalisées parce qu'elles sont associées au statut inferieur des femmes dans ce pays »(Ennaji 2005).Mais quoi qu'il en soit des réalités objectives ou des statuts subjectifs que l'on puisse attribuer aux femmes, à nos mère, à notre langue et culture mère dans notre société, la montée de la société dite connectée nous met en plus devant des défis nouveauxet procède à une redistribution en profondeur des rôles de socialisation et de transmission des savoirs.

3. L'acquisition des savoirs à l'ère numérique

De nos jours, et cela depuis les deux dernières décennies du siècle dernier, avec l'avènement de l'ère numérique et surtout celle du web social (web 2.0) et puis la naissance de ceux qu'on appela dès lors les natifs numériques, le capital culturel de cette nouvelle génération n'a pas cessé de connaitre des bouleversements successifs majeurs. Du coup, les modes et les moyens de socialisation et de l'acquisition des savoirs et des savoir-faire – et aussi des langues - ont basculé vers ce que l'on appela désormais la disposition technologique caractérisée par l'usage abusif des technologies de l'information et de la communication, par la recherche de la facilité et de la rapidité, par les apprentissages et les performances multitâches parallèles et par de nouveaux rapports au temps et à l'espace qui favoriseraient le virtuel aux dépends du monde réel. Je m'arrêterai très brièvement sur quelques caractéristiques du Web 2.0 qui nous intéressent et puis j'adresserais la théorie – ou plutôt l'hypothèse – des natifs numériques en plus de détail.

3.1 L'avènement du Web 2.0

Il semblerait que l'expression Web 2.0 ait fait sa première apparition en 2003 pour designer un ensemble de nouveaux phénomènes qui ont caractérisé le World Wide Web (www), phénomènes qui avaient en commun une dimension plus sociale que jamais de l'interaction sur Internet. Je me contenterai ici de rappeler simplement les caractéristiques phares du Web 2.0[1]. Je citerai en premier lieu que le Web 2.0 considère les utilisateurs comme des entités de première classe dans le système, avec des pages consacrées aux profiles de ces derniers qui comportent leurs informations personnelles comme l'âge, le sexe, la localité, leurs préférences, leurs avis et leurs témoignages ainsi que ceux que peuvent porter d'autres utilisateurs sur eux. La deuxième caractéristique du web social est celle qui permet aux utilisateurs d'établir des connexions entre eux et aussi de former ou de devenir membres dans des groupes et former ainsi des réseaux constitués d' « amis » et de recevoir des nouvelles ou des « mises à jours » de ces amis. La troisième caractéristique du web 2.0 est celle qui

1 Je ne pourrai pas parler des innovations techniques qui ont rendu possible cette nouvelle interaction sociale des utilisateurs sur le net comme les technologies de cryptage, de présentation et de structuration des sites, parce que ces aspects techniques dépassent de loin mes compétences de linguiste.

permet aux utilisateurs de poster des contenus de différents formats : des photos, des vidéos, des blogs, des étiquettes, et de contrôler le « partage » de ces contenus [2]. En résumé, l'enfant qui naît dans le milieu connecté d'aujourd'hui aura bien à sa disposition la possibilité de consulter des contenus aussi variés en formats et pourra s'associer à des groupes en dehors du foyer familial et ne sera jamais limité ni par les contraintes du temps ou de l'espace, ni par la géographie ou les distances. Il est clair que nous sommes bien loin des modes traditionnels de socialisation qui liaient intimement l'enfant à sa mère.

3.2 Les natifs numériques et la disposition technologique

En parallèle avec la monté du Web social, un autre développement, mais cette fois très controversé va se faire dans le discours pédagogique qui soutient que la génération du nouveau millénaire aurait vécu un changement plus profond que laissent entendre les aspects quantitatifs de la dissémination des technologies numériques. En fait selon les auteurs de cette hypothèse (Prensky, M. 2001), la génération du nouveau millénaire aurait vécu une espèce de mutation génétique irréversible et que le cerveau de cette génération penserait et traiterait l'information de façon fondamentalement différente des générations antérieures. Apparemment, les expériences qualitativement nouvelles des générations dites connectées, auraient mené vers une nouvelle structure de notre cerveau. Pour donner un exemple de ces expériences, Prensky (idem) cite le fait que les étudiant en fin de parcours de la licence aujourd'hui auront passé 5000 heures de leurs vie a lire, 10000 heures aux jeux vidéos, et 20000 mille heures à regarder la télévision. En plus, ils auraient manipulé dès leur jeune âge des téléphones portables, des ordinateurs, des jeux vidéo et ils auraient utilisé les emails, la messagerie instantanée de façon instrumentale dans leur vie quotidienne. Par contre, nous qui sommes nés avant les années 90 du siècle dernier, nous serions ce qu'ils appellent les immigrants numériques. Les auteurs de cette hypothèse avancent que nous, les immigrants numériques, serions incapables d'enseigner les natifs numériques parce que, apparemment, nous parlerions un langage obsolète qu'eux ne comprennent pas.

Alors quelles seraient les caractéristiques comportementales de cette génération de soi-disant mutants? Les natifs numériques auraient l'habitude de recevoir l'information très vite. Des qu'une question se pose, la recherche de sa réponse s'impose immédiatement. Comment ne pas le faire ; Ils sont connectés en permanence. Ils aiment aussi que la récompense de leur travail soit instantanée. Deuxièmement les natifs numériques se distinguent par le traitement parallèle d'informations différentes et ils opèrent souvent par des taches multiples à la fois. Les natifs numériques auraient aussi une préférence des graphiques au dépends des textes et ils n'accèdent jamais les contenus de façon linéaire mais plutôt de façon aléatoire un peu comme dans un hypertexte. Et enfin, les natifs numériques n'aiment pas le travail sérieux ; ils préfèrent de loin toutes les activités qui s'apparentent aux jeux. Alors les questions qui s'imposent : doivent les natifs numériques apprendre les bonnes vielles manières, quitte à

2 Une caractéristique du Web 2.0 est plus technique dans le sens où elle permet d'associer des contenus dits « riches » aux contenus disponibles sur la toile tels que les animations flash et la communication avec les autres à travers la messagerie interne des plateformes et les messages instantanés.

savoir par la coercition ? Ou alors, doivent les immigrants numériques – que nous sommes - apprendre les nouvelles manières ? Il semble que la réponse soit évidente. Les natifs numériques ne reviendront jamais en arrière. Et les adultes les plus avisés admettent qu'ils ne savent pas grand-chose du nouveau monde et feront l'effort d'apprendre les nouvelles tendances de leurs petits et d'intégrer ainsi le nouveau monde. Mais, bien que l'hypothèse des natifs numériques ait subit un certain nombre de critiques sévères (voir par exemple, Koutropoulos, A. 2011), je trouve personnellement que s'il ne faut pas la prendre très au sérieux et la généraliser, il n'en demeure pas moins vrai que cette théorie puisse s'appliquer au moins à une certaine souche de la société qui est bien dotée de ces nouveaux moyens technologiques.

4. Les technologies de l'information au Maroc

Passons maintenant à la situation dans notre pays. Est-ce ce que tout cela nous concerne dans une quelconque mesure? Selon le Global Information Technology Report 2016, publié par le Forum Economique Mondial (Silja Baller, S. *et al.* 2016), le Maroc occuperait la 78ème place dans le classement mondial par pays dans la capacité de connexion du réseau national (Network Readiness Index).Le Maroc devance ainsi la Tunisie qui se classe à la 81ème place et loin devant l'Algérie qui serait à la 117ème place. Le Maroc occuperait aussi la 60ème dans le classement des pays en termes de l'exploitation des TIC (Usage Sub-Index) alors que la Tunisie serait à la 80ème place et l'Algérie à la 125ème place. A en croire toujours les statistiques, Il semble qu'en 2016, sur une population estimée à 34 millions d'habitants, plus de 20 millions de Marocains utilisent l'Internet, ce qui représente plus de 60% de l'ensemble des habitants en 2016[3]. Le Maroc vient ainsi en tête de liste en Afrique du Nord devant la Tunisie et l'Algérie ou respectivement plus de 52 pour cent et 37 pour cent de la population utilisent l'internet. Le E-Learning Africa Report 2015 décrit notre pays en termes plus qu'élogieux, je cite : « L'infrastructure TIC du Maroc est l'une des meilleur d'Afrique et souvent prise comme model pour les autres pays. [….] Les connexions fixes sont moins répandues que celles mobiles, mais le gouvernement est en train d'investir dans ce domaine, et à travers le plan national de connexion à haut débit il vise offrir un accès fixe ou mobile à la population entière d'ici 2022 ».

D'autres indicateurs de plus pour notre pays nous font bien rêver. Sur une population de plus de 34 millions d'habitants, dont plus de 66% sont dans la tranche d'âge entre 15 et 60 ans, et avec un taux de scolarisation qui peut atteindre parfois plus de 90 %, le Maroc semble être un candidat idéal pour usage réussi des technologies de l'information et de la communication (TIC) dans l'enseignement et la formation[4]. En effet, les indicateurs pour notre pays ne cessent de s'améliorer. Dans le rapport officiel présenté en 2008 à la conférence internationale sur l'éducation[5], il en ressortait déjà que 15% des écoles primaires, 40% des collèges, et 63% des lycées étaient dotés d'une salle d'informatique ou d'une salle multimédia et que ce matériel informatique serait renforcé annuellement par 5 à 10 ordinateurs pour certains établissements accueillant des classes d'élites.

3 Internet World Statistics 2016 (www.internetworldstats.com)
4 Ministère de l'Education Nationale du Maroc (2008). Rapport National sur Le développement de L'Education.
5 Idem

Le rapport avance le taux global de 50% des élèves qui bénéficiaient d'un cours d'initiation à l'informatique et qu'une plus large couverture reste à prévoir au fur et à mesure que des enseignants soient formés dans les différents centres de formations. Le document précise toutefois que le manque à combler pour mener à bien une initiation de masse aux TIC reste encore très grand. Et si telle est la situation aux niveaux de l'éducation primaire et secondaire, celle de l'enseignement universitaire reste très peu documentée. Sur plus de 322 établissements de l'enseignement supérieur public et 460 établissements de l'enseignement supérieur privé abritant au total environs 400.000 étudiants, aucun indicateur officiel, ou du moins fiable, qui préciserait le taux d'utilisation des TIC dans l'enseignement supérieur n'est actuellement disponible. Et en ce qui concerne le recours au mode e-learning, la situation semble être moins claire. Bien que l'ensemble des établissements marocains de l'enseignement supérieur dispose en principe d'un portail d'enseignement à distance, l'enseignement en ligne dans ces établissements n'est utilisé que comme support de cours à caractère facultatif. Le E-Learning Africa Report 2015, d'ailleurs note que: « Le Maroc est critiqué pour son manque en e-learning en dépit des conditions excellentes pour son implémentation ». Le rapport cite, toutefois, quelques initiatives réussie mais qui restent en deçà du potentiel de notre pays[6]. Nous sommes donc en mesure de nous demander si oui ou non les indicateurs positifs des usages des TIC ont un quelconque impact réel sur la qualité et la modernité de l'enseignement en général dans notre pays.

5. La génération connectée et le paradoxe

Mais en dépit du manque en initiatives efficaces d'enseignement via les TIC, nos jeunes affichent une toute autre réalité. En effet, n'est-il pas frappant de voir ces jeunes technophiles dans nos écoles et dans nos lycées, qui ne se séparent que très rarement de leur téléphone portable dernière génération ou ceux sur les terrasses des cafés ou dans nos campus qui sont tout le temps collés à leur téléphone portable dernier cri ou leur PC portable. Est-ce que ces jeunes sont tous des « adopteurs précoces » des nouvelles technologies ou forment-ils une certaine « majorité avancée »? Selon une typologie bien connue du cycle de diffusion et d'adoption des technologies de l'information(Moore, G. 1998), nos jeunes technophiles marocains ne seraient peut-être que des « adopteurs précoces » toujours avides d'acquérir les dernières innovations technologiques pour satisfaire un orgueil personnel, et il ne serait pas certain non plus qu'ils soient suivis par une quelconque majorité avancée qui associe la nouveauté à l'efficacité et à la productivité[7]. Et bien qu'il soit tentant d'opter pour des généralisations aussi simplistes, la réalité est plus complexe qu'elle semble à priori l'être. D'autres faits nous apportent des éléments de réponse à ces questions bien différentes de telles conclusions simplistes.

6 Le rapport cite quand même des initiatives comme la communauté de bienfaisance Dar Si Hmad qui organise des cours à distance pour préparer les élèves au milieu rural aux examens et le projet Itqane qui vise former les enseignants dans les régions Fes-Boulmane et Doukkala-Abda à faire face au problème de déperdition scolaire qui serait de l'ordre de 250 000 à 300000 élèves par an.

7 Geoffrey Moore (1995) pour rappel identifie cinq phases par les quelles passe toute innovation technologique. Chacune de ces cinq phases est associée à une catégorie d'utilisateurs : la phase des innovateurs (2,25%), celle des passionnés qu'il appelle *adopteurs précoces* (15%), la phase de la *majorité avancée* (34%), celle de la *majorité tardive* (34%) et, en dernier lieu, la phase des *retardataires* (15 %).

6. Des Eléments de réponse

Il est clair qu'au Maroc, le manque en infrastructure n'est peut-être pas aussi grave que le manque en formation aussi bien des apprenants que des enseignants et ceci ne permet guère une mise en marche fructueuse d'un enseignement assisté par l'outil numérique. Il est intéressant quand même de noter comme le démontre une enquête réalisée à deux reprises, en 2007 et en 2010 a Taza, (El Halfaoui, M. 2013) que 84% des enseignants universitaires sont plutôt favorables à l'introduction des TIC dans leurs enseignements. Mais, la quasi-totalité des sujets étudiés (100%) dans cette enquête exprime un niveau très élevé d'anxiété face aux environnements numériques, et entre 74% et 81% disent avoirs des compétences modestes dans l'utilisation des outils numériques. Donc deux facteurs font obstruction à une exploitation du numérique dans l'enseignement au Maroc, les insuffisances en infrastructure et les insuffisances en moyens humains qualifiés. Ironiquement, et d'après une étude très bien documentée que nous offre M. T. Saliba (2013), même dans un pays comme le Liban où 96% des élèves possèdent un ordinateur qu'ils utilisent entre 2 et 10 heures ou plus par semaine, et même quand ces élèves affichent une grande satisfaction quand au rôle joué par l'école dans leur initiation aux différentes applications informatiques, il apparaît que 75% de ces élèves n'utilisent leurs ordinateurs que pour le «chat» ou pour consulter des réseaux sociaux comme Facebook. Seulement 25% des sujets dans l'étude en question admettent que l'informatique enseignée à l'école joue un rôle important dans leur préparation au monde du travail et de la production contre 32% de ces mêmes sujets qui pensent que l'initiation à l'outil informatique et aux technologies de l'information dispensée par l'école favorise d'abords leur utilisation comme des « moyens de communication humaine ».

Cette dernière conclusion nous interpelle et nous rappelle que les technologies, numériques ou autres, ne sont bonnes que si elles répondent à un besoin d'interaction humain. Rappelons-nous au moins une des caractéristiques fondamentales du Web social : les utilisateurs sont des entités de première classe dans le système. Donc, les technologies qui ne mettent pas la dimension humaine et sociale des utilisateurs en premier lieu n'auront pas de succès à long terme. Prenons par exemple les cours ouverts massifs en ligne, connus sous le nom de MOOC. Un certain nombre de critiques notent que le grand nombre d'individus qui peuvent suivre un MOOC n'estpas le seul critère déterminant de succès. Ces critiques notent aussi le taux de décrochage aux MOOCs aujourd'hui qui dépasserait les 80% (Prensky, M. 2013).Et à l'opposé des MOOCs qui nécessitent des plateformes particulières et donc une infrastructure coûteuse, le succès de la messagerie téléphonique, le SMS, nous laisse émerveillés. Comme sous-produit de la téléphonie mobile, cette technologie pourtant simple n'a pas cessé de grandir depuis 25 ans environ. Il semble que plus de 4 milliards de personnes échangent des SMS à longueur de journée maintenant. Apparemment, le SMS est en train de changer même les habitudes de notre rédaction et notre écriture ; il est en train de remodeler la langue de demain. Un autre outil miracle qui est souvent cité dans les études de développement linguistique assisté par ordinateur n'est autre que le traitement de texte (Motteram, G. 2013). En fait, la théorie même selon la quelle l'initiation à la rédactionserait plus complexe et passerait par quatre phases, celle de la pré-rédaction, la rédaction, la révision et l'édition (Tribble, C. 1996) est en fait une parfaite correspondance entre la théorie de l'apprentissage

et le modèle informatique du traitement de texte. Dans tous les cas, d'après les enquêtes de terrain (Pennington, MC. 1996), le traitement de texte nous permet d'écrire plus facilement, d'écrire plus, d'écrire différemment, et d'écrire mieux.

L'histoire de l'écriture nous enseigne aussi que les outils utilisés pour écrire ont bien façonné l'art et la manière d'écrire. Un exemple frappant qui est souvent cité (Kern, R. cité parMotteram, G. 2013) est celui du chinois qui parce qu'il était écrit sur des bande extraite des cannes de bambou en est arrive à être écrit en colonnes verticales durant des siècles et des siècles. Et s'il ne fallait parler que de typographie, sans aller bien loin, de nos jours nous disposons de puissants programmes informatisés de saisi, de visualisation et d'impression des textes qui possèdent des fonctionnalités typographiques programmables qui dépassent de loin tout ce que nous pouvions imaginer et écrire manuellement par le passé. La production et l'édition de différentes versions, différentes familles et différents types de polices est très rapide, plus rapide que ne l'a jamais été auparavant (Noordzij, G. 2006). Ces outils qui servent à développer l'écrit sont d'une importance primordiale dans le cas particulier du Tamazight. Pour une langue qui a en plus le handicap de posséder un alphabet qui a cessé d'évoluer depuis au moins des siècles, développer l'art et la manière de l'écrire semble lui ouvrir les voies du future.

Et enfin, rappelons-nous aussi une autre caractéristique de ces générations futures que nous appelons les natifs numériques :ils préfèrent de loin toutes les activités qui s'apparentent aux jeux. Alors tous les outils à caractère ludique ne peuvent qu'être populaires. Autres que les jeux vidéo, les systèmes destinés à générer des documents multimédias comprenant une synchronisation de la parole et de l'écrit ne peuvent qu'avoir le succès qu'ils méritent. Ces systèmes qui s'apparentent au karaoké s'avèrent merveilleux pour l'association de l'oral à l'écrit. En fait, nous avons tous eu le plaisir un jour d'assister ou même prendre part à une soirée de Karaoké, alors faisons de même pour les contenus que nous voudrions enseigner pour une meilleur association de l'utile à l'agréable.

7. Conclusion

En guise de conclusion, il est bien évident que les développements technologiques de l'ère numériques représentent des défis de grande taille pour les sociétés dites à faible potentiel technologique comme la nôtre. Ces défis sont beaucoup plus ardus pour le Tamazight qui reste une des langues les moins dotée des nouvelles TIC au monde. Si j'ai insisté sur les outils qui servent à développer l'écrit et l'association de l'écrit à l'oral, c'est parce que dans le cas particulier du Tamazight l'écrit semble présenter des difficultés singulières. Mais heureusement que les technologies modernes offrent aussi des opportunités inégalées, et autrefois inimaginables, pour une plus large diffusion et donc une revitalisation de cette langue. Il est temps que le Tamazight noue des liens étroits avec le medium numérique dans sa dimension sociale adapté aux besoins des générations futures. Certaines technologies réussiront plus que d'autres, mais la langue aura servi à engager des utilisateurs dans des activités et des usages qui laisseront une trace profonde dans la langue elle-même. Notons enfin que les technologies elles-mêmes passent par des cycles obligatoires de vie. D'abords elles sont restreintes à un public de spécialistes. Puis elles sont ouvertes sur le grand public. Et enfin, après un certain temps, elles sont tellement intégrées que nous n'arrivons plus à

les remarquer en tant que technologies « nouvelles ». Elles deviennent normalisées comme le sont maintenant le tableau, la craie et l'ardoise (Bax, S. 2011).Alors, dans un monde complètement envahit par les téléphones, les ordinateurs, et les autres gadgets, il serait temps peut-être pour que le numérique et le Tamazight s'intègrent et que la fracture entre les deux disparaisse.

Références

Bax, S. (2011). Normalisation revisited: The effective use of technology in language education. *IJCALLT* 1/2: 1–15.

Becker, Cynthia. (2006). Amazigh Textiles and Dress in Morocco: Metaphors of Motherhood. *African Arts*. Vol. 39, No. 3, Pages 42-96. MIT Press.

Bourdieu, Pierre (1979). « Les trois états du capital culturel. » In: Actes de la recherche en sciences sociales. Vol. 30, novembre 1979. L'institutionscolaire. pp. 3-6.

Ennaji, Moha. (2005). *Multilingualism, Cultural Identity, and Education in Morocco*. New York: Springer.

Fishman, Joshua A. & Garcia, Ofelia.Eds. (2010). *Handbook of Ethnic Identity: Disciplinary and Regional Perspectives*. Oxford University Press.

Gelb, Ignace (1963). *A Study of Writing*. Chicago: University of Chicago Press.

Halfaoui, M. (2013). « Pratiques et profils d'utilisation des TICE: Les enseignants d'une faculté ». *In L'environnement numérique dans l'ensignement des langues*. (Edité par IdrissiAydi, O. et Marjane, D.) Maison d'édition post-modernité, Fès (2013).

Koutropoulos, Apostolos. (2011). «Digital Natives: Ten Years After». *MERLOT Journal of Online Learning and Teaching*.Vol. 7, No. 4: 525-538.

Manji, F. E. Jal, Badisang, B. Opoku-Mensah, A. (2015). The E-Learning Africa Report: *The Trajectory of Change*. ICWE GmbH Leibnizstr. Berlin: Germany

Moore, Geoffrey A. (1995). *Inside the Tornado: Marketing Strategies from Silicon Valley›s Cutting Edge*. HaperCollins.

Motteram, G. Ed. (2013). "Developing and extending our understanding of language learning and technology." *In Innovations in Learning Technologies for English Language Teaching* (Ed. G. Motteram) London: British Council

Noordzij, G. (2006). The Stroke: Theory of Writing. Princeton Architectural Press.

Pennington, MC (1996) *The power of CALL*. Houston: Athelstan.

Prensky, Marc (2001). « Digital Natives, Digital Immigrants », *On the Horizon*: Vol. 9, No. 5: 1-6.

Sadiqi, Fatima (2003). *Women, Gender and Language in Morocco*. Brill: Leiden-Boston. Springer.

Saliba, M. T. (2013). « L'intégration de l'Informatique et de la Technologie en disciplines scolairesdans les écoles privées au Liban ». *In L'environnement numérique dans l'ensignement des langues*. (Edité par Idrissi Aydi, O. et Marjane, D.) Maison d'édition post-modernité, Fès (2013).

Silja Baller, S. Dutta& B. Lanvin. Eds. (2016). Global Information Technology Report 2016. Innovating in the Digital Economy. The World Economic Forum Geneva and INSEAD.

Tapscott, Don (1998). *Growing Up Digital: The Rise of the Net Generation*. New York: McGraw & Hill.

Tribble, C. (1996) *Writing*. Oxford: Oxford University Press.

Weber, Max. (1968). *Economy and society: An outline of interpretive sociology*. New York: BedminsterPress.

Analyse Syntaxique Automatique en Tamazight (Kabyle)

Farida Yamouni [1]
[1] Université Mouloud Mammeri Tizi Ouzou Algérie
fariyamo@yahoo.fr

Résumé

Dans cet article nous présentons la description des grammaires syntaxiques pour la phrase simple en Tamazight (kabyle) en vue de l'analyse syntaxique automatique de phrases. Nous retenons la phrase simple déclarative, avec la phrase simple nominale et la phrase simple verbale. Nos travaux entrent dans le cadre de développement du module Tamazight (kabyle) dans la plate-forme NooJ[1] . Nous élaborons les ressources linguistiques utilisées pour l'analyse syntaxique et nous décrivons les mots grammaticaux recensés et utilisés, autres que le nom et le verbe. Nous présentons ensuite les grammaires syntaxiques construites ainsi que des exemples d'analyse syntaxique de phrases en Kabyle. L'analyse linguistique qui est tout d'abord effectuée avant l'analyse syntaxique utilise les dictionnaires électroniques existants (Aoughlis *et al.*, 2013), (Aoughlis, 2014) afin de fournir les catégories syntaxiques utilisées lors de l'analyse syntaxique. Les grammaires syntaxiques permettant l'analyse sont construites sur la base des travaux de Nait-Zerrad K. à l'aide du système NooJ en vue du traitement automatique de la langue Tamazight (kabyle).

1. Introduction

Une implémentation des règles de grammaire est réalisée dans NooJ en vue de l'analyse syntaxique automatique de phrases. Nous nous intéressons à la phrase simple déclarative. Nous décrivons la phrase simple avec ses constituants. Elle est composée d'un énoncé minimal : EVM (Enoncé Verbal Minimum) ou bien ENM (Enoncé Nominal Minimum), suivi d'un ou plusieurs compléments (GN : groupe nominal). La phrase verbale complète est l'EVM. La phrase nominale complète est l'ENM. Une étude de la grammaire de Tamazight (Kabyle) est faite. Les grammaires syntaxiques sont construites pour les deux types de phrases simples. Des tests sont réalisés dans NooJ pour analyser automatiquement des phrases de la langue retenue. Les résultats sont présentés pour quelques phrases.

2. La phrase simple

La phrase met en relation deux éléments : un sujet (ce dont on parle) et un prédicat (ce que l'on dit du sujet), suivis éventuellement de compléments. Toute phrase est constituée d'un

1 http://www.nooj4nlp.net/

prédicat complété par au moins un groupe nominal qui lui est lié. Le prédicat peut être verbal ou nominal.

La phrase simple est constituée d'un EVM ou d'un ENM suivi d'un ou plusieurs compléments. Selon (Chaker, 1998), « en berbère, langue à opposition verbo-nominale, c'est le verbe qui constitue généralement le noyau prédicatif, dans certains contextes, notamment en proposition relative, le verbe peut perdre cette fonction et devenir un déterminant lexical équivalent à un adjectif ».

Le nom peut occuper la fonction de prédicat. Les fonctions du nom sont (Chaker, 1998 ; Nait-Zerrad, 2001):

Le CR, Complément Référentiel ou explicatif : il est marqué par l'état d'annexion, postposé à l'élément qu'il détermine, souvent le prédicat ;

Le COD, Complément d'Objet Direct : c'est le nominal directement placé soit après le verbe, soit après le CR, il est à l'état libre. Il peut être remplacé par un pronom affixe complément direct ;

Le COIND, Complément d'Objet INDirect, est obligatoirement relié au verbe par la préposition «i», à. Le noyau nominal est à l'état d'annexion ;

L'IT, Indicateur de Thème : le syntagme nominal à l'état libre se trouve en position initiale, c'est une mise en relief pour annoncer l'indice de personne ou le pronom personnel ;

Le CC, Complément Circonstanciel : il est introduit le plus souvent par une préposition. On peut avoir plusieurs CC se rapportant à un même verbe ou équivalent nominal. On trouve le complément direct et les compléments prépositionnels (CC de lieu, cause, but, manière, moyen, accompagnement, comparaison) ;

Le CD, Complément du nom (nom déterminant un autre nom) ou Complément Déterminatif, dans ce cas là le nom a une relation indirecte avec le verbe. Il peut suivre directement le nom déterminé ou être précédé de la préposition «n», de. Il est toujours à l'état d'annexion ;

L'Adjectif ;

Le Participe : le participe est un verbe pour lequel les oppositions de personnes sont neutralisées, il s'agit donc de constructions de type relatif dans lesquelles le verbe a perdu sa fonction de prédicat.

2.1. La phrase simple verbale

Elle est constituée d'un EVM (énoncé verbal minimum) suivi d'un ou de plusieurs compléments : GN ayant la fonction de COD, COIND, CR, CC, etc. (Nait-Zerrad, 2001 ; 1995b).

L'EVM est composé d'un verbe + indice de personne.

La structure de la phrase simple verbale est :

EVM/CR(EA)/COD(EL)/COIND/[CC]* ou bien :

EVM (+pCOIND+pCOD+Po)/CR/CR1(COD repris)(EA)/COIND/[CC]

Avec :

COIND= PREP «i» + (N+EA) ;

Po= particule d'orientation ;

pCOD= pronom personnel COD ;

pCOIND= pronom personnel COIND ;

CR1= 2ème CR correspondant au COD, car il reprend un pronom COD.

2.2. La phrase simple non verbale

Elle est constituée d'un EVM (énoncé verbal minimum) suivi d'un ou de plusieurs compléments : GN ayant la fonction de COD, COIND, CR, CC, etc. (Nait-Zerrad, 2001 ; 1995b).

Elle est aussi appelée phrase nominale. C'est une phrase sans verbe, constituée d'un ENM, suivi d'un ou plusieurs compléments ou GN : CR, IT, CC, adjectif, etc.

Phr-Simpl-Nom = ENM + GN*

L'ENM contient un nom ou équivalent (jouant le rôle du verbe dans l'EVM), accompagné d'un élément prédicatif. L'élément prédicatif peut être :

- La particule «d» dite prédicative, placée devant le nom à l'état libre ;
- Une préposition ;
- Un présentatif.

2.2.1. L'ENM avec la particule prédicative «d» :

L'ENM est composé entre autres de :

d + nom : d aqcic ;

d + adjectif : d ameqqran ;

d + pronom : d nutni ;

2.2.2. L'ENM avec la préposition «n»

Elle indique l'appartenance, nous avons entre autres :

n + pronom ;

n + nom : n mmi-s, userwal-agi (CR); Aserwal-agi (IT), n mmi-s ;

n + démonstratif : aγrum, n ta ;

2.2.3. La phrase simple nominale

Nous avons les exemples suivants :

- Pronom + ENM : Netta d aqcic. / Lui, c'est un garçon ;
- Nom + ENM : Axxam d ameqqran / La maison (elle) est grande.

3. Les ressources linguistiques pour l'analyse syntaxique

Dans le système NooJ (Silberztein, 2003), les grammaires syntaxiques sont représentées par des ensembles structurés de graphes ou de règles qui décrivent des paradigmes syntaxiques stockés dans des fichiers «.nog».

Pour l'analyse syntaxique automatique, nous avons construit des grammaires syntaxiques pour l'analyse de la phrase simple verbale et de la phrase simple nominale.

Une analyse linguistique est d'abord faite pour récupérer les catégories syntaxiques obtenues grâce à la consultation des dictionnaires élaborés dans NooJ pour le nom et le verbe. Ces catégories sont utilisées pour l'analyse syntaxique. Nous avons aussi recensé les mots grammaticaux et complété le dictionnaire existant.

3.1. Les dictionnaires utilisés

Plusieurs dictionnaires sont utilisés pour l'analyse linguistique, ils permettent de récupérer les catégories syntaxiques utilisées pour l'analyse syntaxique. Le dictionnaire des noms et celui des verbes sont décrits dans (Aoughlis et al, 2013). Le dictionnaire des verbes simples contient 3900 entrées, celui des noms 500 entrées. Les noms et adjectifs qui ont été dérivés des verbes (Aoughlis, 2014) sont dans le dictionnaire des noms et adjectifs dérivés.

3.2. Le dictionnaire des mots grammaticaux

Nous avons recensé et codé les catégories syntaxiques (Nait-Zerrad, 1995a) autres que le nom et le verbe, dans le dictionnaire des mots grammaticaux qui contient 116 entrées. Nous trouvons ci-dessous, un extrait :

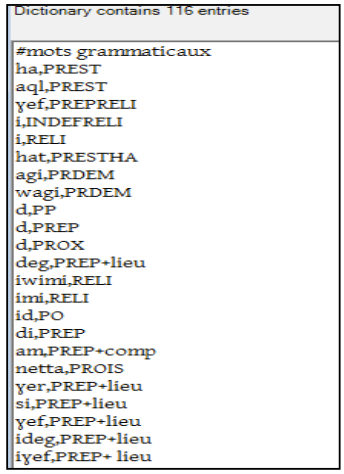


Figure 1. Extrait de la grammaire des mots grammaticaux

3.3. Les grammaires syntaxiques

Deux grammaires principales sont construites : une pour la phrase simple verbale et une autre pour la phrase simple nominale. Les grammaires sont utilisées pour l'analyse syntaxique dans NooJ.

3.3.1. Grammaire syntaxique de la phrase simple nominale

Elle contient 9 graphes. Nous trouvons ci-dessous la grammaire principale PHR-SIMPL-NOM.nog :

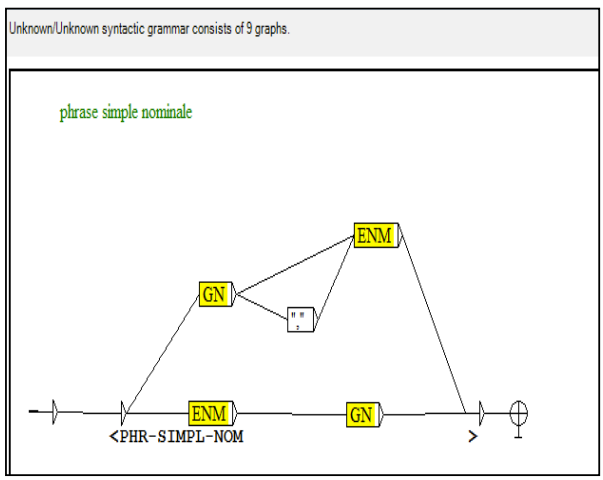


Figure 2. Extrait de la grammaire PHR-SIMPL-NOM

PHR-SIMPL-NOM est composée d'un ENM suivi d'un GN ou bien d'un GN suivi d'un ENM. Nous avons deux sous-graphes ENM et GN. Nous trouvons ci-dessous, le sous-graphe ENM, avec les règles de grammaires. Quatre sous-graphes sont utilisés : REL pour la relative, ENMn pour l'ENM avec «n», PRESTAHA et PRESTAQL.

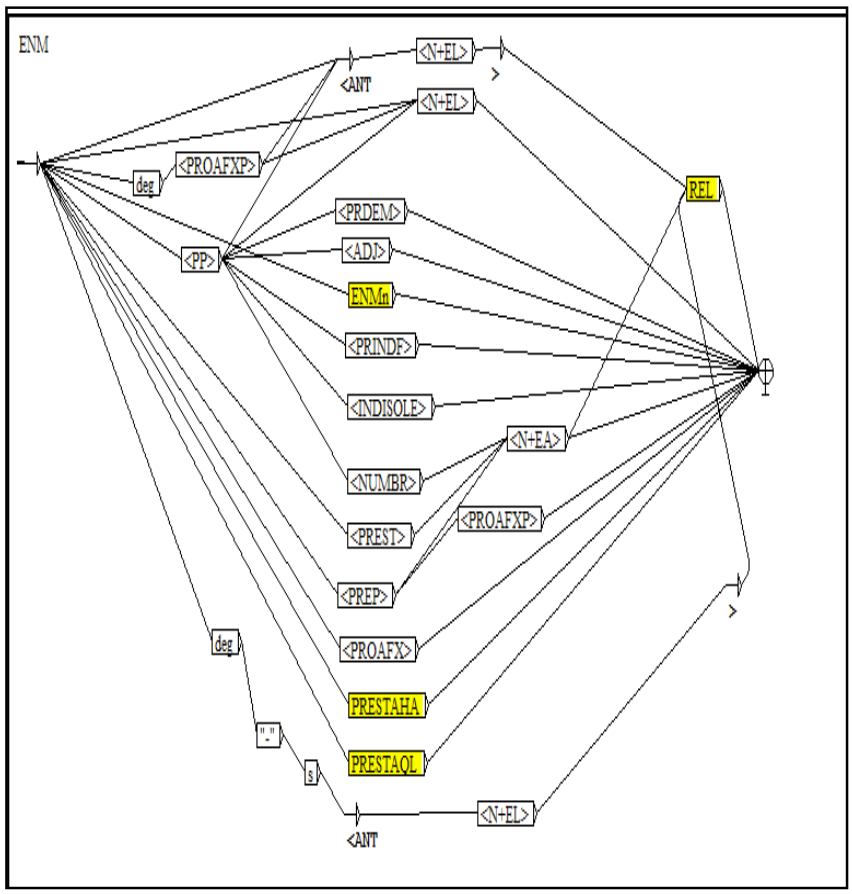


Figure 3. Le sous-graphe ENM

3.3.2. Grammaire syntaxique de la phrase simple verbale

Nous trouvons ci-dessous la grammaire principale PHR-SIMPL-VERB, avec par exemple, les sous-graphes EVM, GNL (sous-graphe décrit dans la figure 5 suivante) ou GN état libre, GNA ou GN état d'annexion.

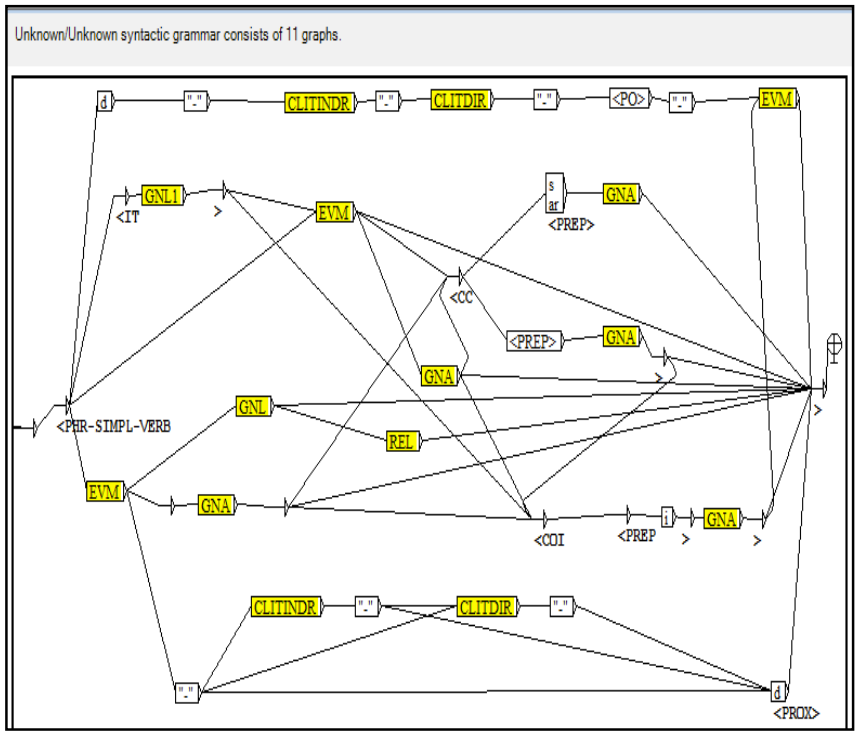


Figure 4. Extrait de la grammaire PHR-SIMPL-VERB

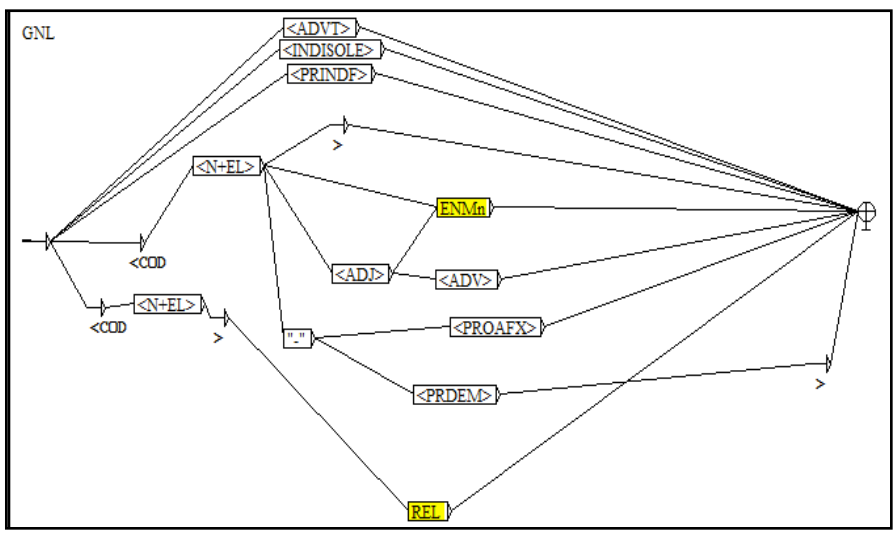


Figure 5. Le sous-graphe GNL

4. Analyse syntaxique de phrases

4.1. Analyse syntaxique d'une phrase simple nominale

Soit la phrase à analyser : Axxam d ameqqran. / La maison (elle) est grande

La grammaire utilisée est PHR-SIMPL-NOM.nog

Résultats de l'analyse :

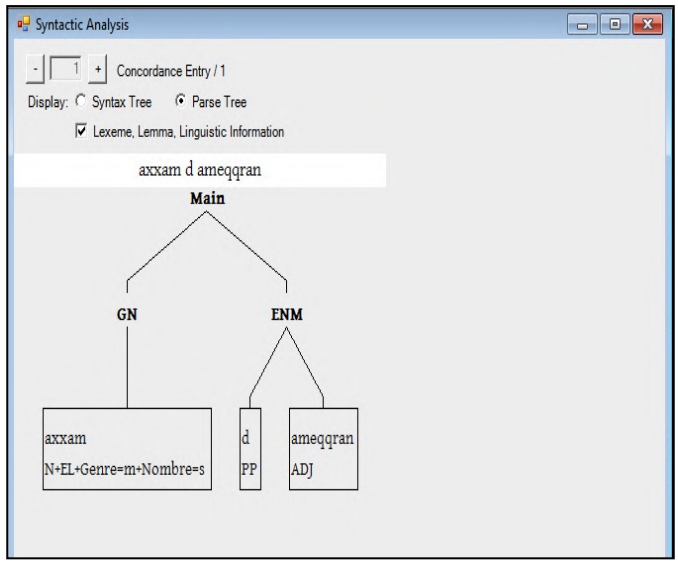


Figure 6. Arbre d'analyse de la phrase «axxam d ameqqran»

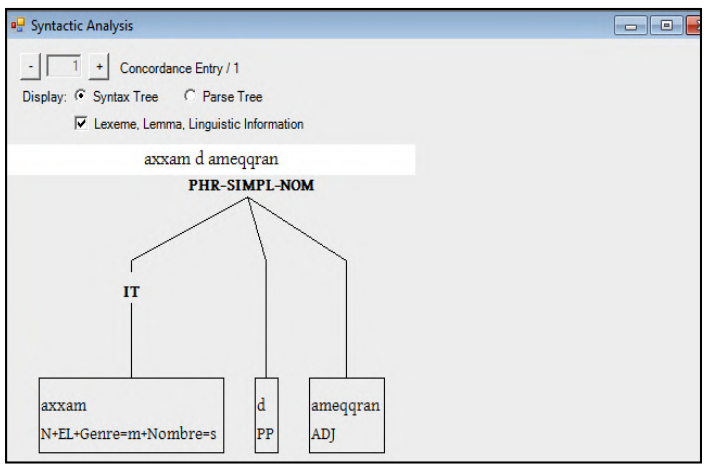


Figure 7. Arbre syntaxique de la phrase»axxam d ameqqran»

Soit une autre phrase à analyser : amcic-agi n yemma, les résultats sont :

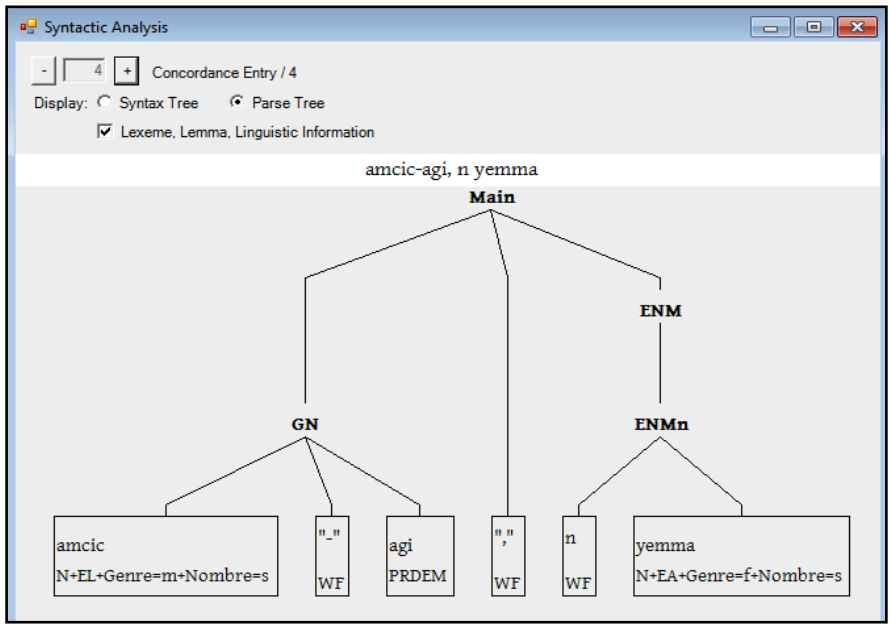


Figure 8. Arbre d'analyse de la phrase «amcic-agi n yemma»

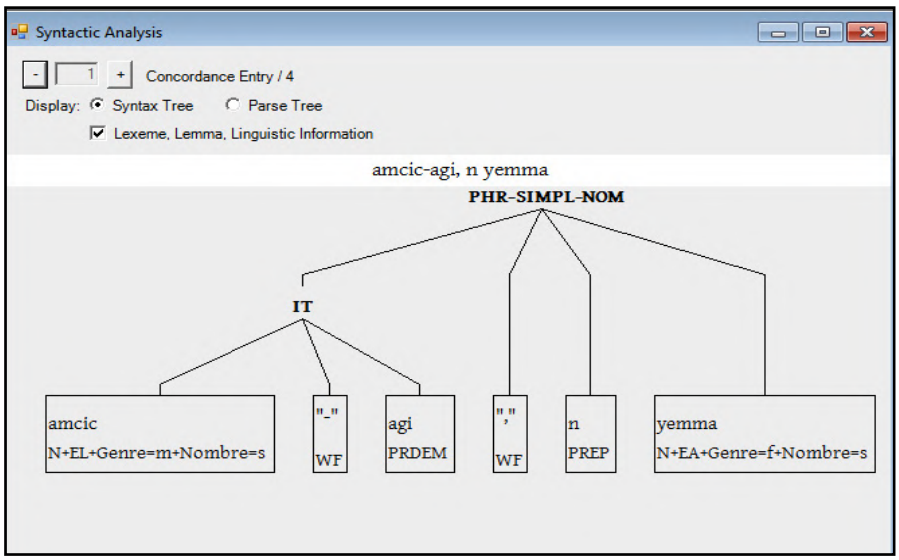


Figure 9. Arbre syntaxique de la phrase»amcic-agi n yemma»

4.2. Analyse syntaxique d'une phrase simple verbale

Soit la phrase à analyser : yekcem wergaz yeččan aγrum ;

la grammaire est PHR-SIMPL-VERB, les résultats des analyses sont :

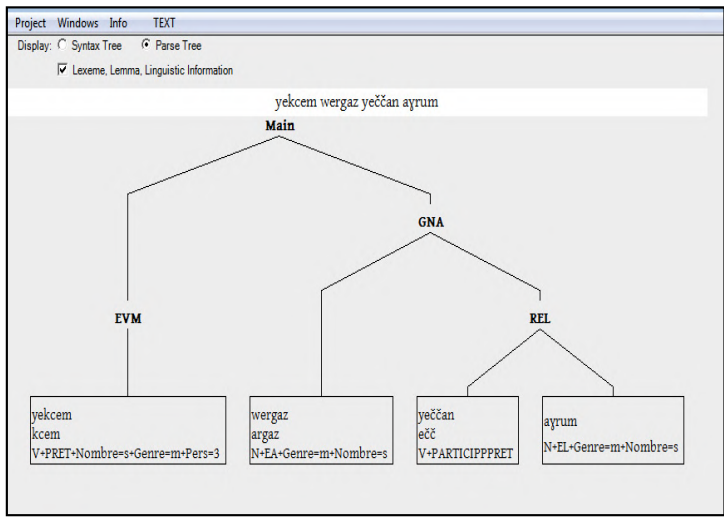


Figure 10. Arbre d'analyse de la phrase «yekcem wergaz yeččan aγrum»

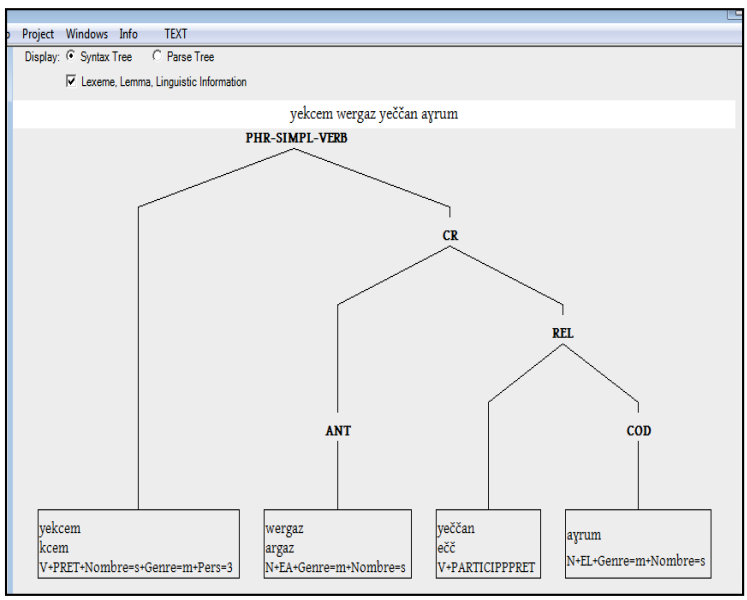


Figure 11. Arbre syntaxique de la phrase «yekcem wergaz yeččan aγrum»

5. Conclusion et perspectives

Dans ce travail nous avons étudié les règles de grammaire du Kabyle, nous avons construit les ressources linguistiques pour l'analyse syntaxique automatique. Les travaux réalisés permettent d'analyser les phrases simple et nominale. Les résultats souhaités sont obtenus et les arbres d'analyse et syntaxiques sont affichés. Les grammaires développées utilisent les catégories syntaxiques définies pour les mots grammaticaux de la langue, ainsi que le nom et le verbe.

Il est envisagé une étude détaillée de la proposition relative avec la mise en place des grammaires syntaxiques. Une étude de la phrase négative est prévue.

Les résultats obtenus peuvent être utilisés pour concevoir et développer un outil pédagogique pour enseigner la grammaire du Kabyle.

Références

Aoughlis F., Nait-Zerrad K., Annouz H., Aït-Kaci F. and Habet MS. (2013). A new Tamazight Module for NooJ, In *Formalising Natural Languages with NooJ 2013*. Selected Papers from the NooJ 2013 International Conference. Svetla Koeva, Slim Mesfar, Max Silberztein Eds. Cambridge Scholars Publishing: Newcastle upon Tyne.

Aoughlis ,F. (2014). Noms et adjectifs dérivés en Tamazight, *6ème Edition de la Conférence Internationale sur Les Technologies d'Information et de Communication pour l'Amazighe,TICAM'14,* IRCAM, Rabat, 24 et 25 novembre 2014. (Actes à paraître aux éditions de l'IRCAM).

Chaker S. (1998). Fonctions, in *Filage-Gastel, Encyclopédie Berbère 19* : 2880-2886. Edisud, Aix-en-Provence.

Nait-Zerrad, K. (1995a). *Grammaire du berbère contemporain (kabyle)*, I-Morphologie , Enag Ed. : Alger

Nait-Zerrad, K. (1995b). *Grammaire du berbère contemporain (kabyle)*, II- Syntaxe, Enag Ed., Alger.

Nait-Zerrad, K. (2001). *Grammaire moderne du kabyle, tajerrumt tatrart n teqbaylit,* Karthala, Paris.

Silberztein, M. (2003). Manuel NOOJ. http://www.nooj4nlp.net

Vers un module d'analyse morphologique de la langue amazighe en utilisant la plateforme NooJ

Fatima Zahra Nejme[1] Siham Boulaknadel[2] Driss Aboutajdine[1]
[1] LRIT, Unité Associée au CNRST (URAC 29), Faculté des Sciences,
Mohammed V-Agdal, Rabat, Maroc.
Fatimazahra.nejme@gmail.com aboutaj@fsr.ac.ma
[2] IRCAM, Avenue Allal El Fassi, Madinat Al Irfane, Rabat-Instituts, Maroc
Boulaknadel@ircam.ma

Résumé

L'analyse morphologique est une tâche primordiale dans divers domaines du traitement automatique des langues et constitues un élément essentiel et une base élémentaire pour bien mener des recherches linguistiques. Conscientes de ce fait et dans la perspective de doter la langue amazighe d'outils et de ressources linguistiques nécessaires pour son traitement automatique, nous avons entrepris de développer un système de construction du lexique et d'analyse morphologique pour l'amazighe. Ce système profite des apports des modèles à états finis au sein de l'environnement linguistique de développement NooJ en faisant appel à des règles grammaticales à large couverture.

1. Introduction

Le Maroc, à l'instar des autres pays du Maghreb, présente le principal état amazighophone avec environ la moitié de la population totale (Boukous, 1995). Cette langue est considérée comme un constituant éminent de la culture marocaine et ce par sa richesse et son originalité. Cependant, il a été longtemps écarté sinon négligé en tant que source d'enrichissement culturel malgré son usage important. Toutefois, et depuis quelques années, la langue amazighe a jouit d'un statut institutionnel par la création de l'institution gouvernementale de la Culture Amazighe (IRCAM- Institue Royal de la Culture Amazighe). Cette création a permis à cette langue ainsi qu'à sa culture de retrouver leur place légitime dans de nombreux domaines. Par conséquent, les premières tentatives de normalisation pour l'apprentissage de cette langue ont étaient amorcées. La démarche d'élaboration a été initiée par la construction des lexiques (Kamel, 2006; Ameur *et al.*, 2009a), l'homogénéisation de l'orthographe et la mise en place des règles de segmentation de la chaîne parlée (Ameur *et al.*, 2006), et par l'élaboration des règles de grammaire (Boukhris et. al., 2008).

Néanmoins, toutes ces étapes sont élémentaires et insuffisante pour qu'une langue peu dotée informatiquement, telle que l'amazighe, puisse franchir le seuil de la mondialisation informatique et de rejoindre ses consœurs dans le domaine des nouvelles technologies de l'information et de la communication (NTIC). Dans ce contexte, de nombreuses recherches

scientifiques sont lancées afin d'améliorer la situation. Ces recherches peuvent être divisées en deux catégories : (1) la construction des ressources computationnels (Es Saady *et al.*,2010; Fakir *et al.*, 2009 ; Boulaknadel *et al.*, 2011), et (2) celle s'occupons de la conception des ressources linguistiques et outils de traitement automatique (Taghbalout *et al.*, 2015 ; Talha *et al.*, 2015 ; Nejme *et al.*, 2013 ; Nejme *et al.* 2015 ; Nejme *et al.* 2016 ; Ataa Allah and Boulaknadel, 2010; Outahajala *et al.*, 2010). Pour notre part, et afin de participer à ce mouvement, nous avons entrepris d'implémenter une composante d'analyse morphologique des textes amazighe en utilisant les apports des modèles à états finis au sein de l'environnement linguistique de développement NooJ en faisant appel à des règles grammaticales à large couverture. Cette composante servira à mieux comprendre la langue partant d'une description de son vocabulaire et de sa morphologie.

2. Description de la langue amazighe

La langue amazighe connue aussi par le berbère ou Tamazight (ⵜⴰⵎⴰⵣⵉⵖⵜ), est une branche de la famille de langue afro-asiatique (chamito-sémitique) (Greenberg, 1966; Ouakrim, 1995). Elle présente la langue émanent d'une des plus anciennes civilisations de l'humanité et couvrent toute l'Afrique du nord, le Sahara et une partie du Sahel ouest africain. Au Maroc, l'amazighe se répartit selon trois grandes zones régionales couvrent l'ensemble des régions montagneuses : le Tarifit au Nord, le Tamazight au Maroc central et au Sud-Est et le Tashelhit au Sud-Ouest et dans le Haut-Atlas. Chacun de ces dialectes évoluent de manière séparée et constituent de plus en plus des communautés sociolinguistiques distinctes et isolées.

Jusqu'à 1994, l'amazighe a été exclusivement réservée au domaine familial (Boukous, 1995). Cependant, et suite à la création de l'IRCAM, l'amazighe a jouit auprès de sa consœur l'arabe d'un statut d'une langue officielle.

La langue amazighe présente une morphologie riche et une structure complexe. Elle présente une langue fortement flexionnelle et montre également la dérivation à haut degré. Afin de couvrir la morphologie amazighe, quatre catégories lexicales doivent être étudiées: les noms, les verbes, les pronoms et les mots outils.

a. Les noms

Le nom est une unité lexicale formée d'une racine et d'un schème. Il possède deux caractéristiques, la première est qu'il peut prendre différentes formes à savoir: une forme simple (ⴰⵔⴳⴰⵣ [argaz] "homme"), forme composée (ⴱⵓⵀⵢⵓⴼ [buhyyuf] "la famine") ou bien forme dérivée (ⴰⵎⵙⴰⵡⴰⴹ [amsawaḍ] "la communication"). La deuxième caractéristique correspond à la variation : il varie en genre (féminin, masculin), en nombre (singulier, pluriel) et en état (libre, annexion).

b. Les verbes

Le verbe peut apparaître sous une forme simple ou dérivée. La forme simple est composée d'une racine et d'un radical. Alors que la forme dérivée est obtenue à partir des verbes simples par la préfixation d'un morphème ou d'une combinaison de deux morphèmes de

valeur différente. Ces morphèmes sont : ⵙ [s]/ⵙⵙ [ss] qui indique la forme factitive, ⵜⵜ [tt] qui marque la forme passive et ⵎ [m]/ⵎⵎ [mm] qui désigne la forme réciproque.

Le verbe, qu'il soit simple ou dérivé, se conjugue selon quatre thèmes: l'aoriste, l'inaccompli, l'accompli positif et l'accompli négatif, et possède trois modes : l'indicatif, l'impérative et participiale.

c. Les pronoms

Le pronom amazighe correspond à tout élément qui peut représenter un groupe nominale ou désigner une personne qui participe à la communication. Les paradigmes des pronoms comprennent : les pronoms personnels, personnels, démonstratifs, interrogatifs et indéfinis. La flexion pronominale est présentée pour les trois catégories : possessifs, démonstratifs et interrogatifs (ⵡⵉⵏⵏⵙ [winns] "le sien" ⵡⵉⵏⵏⵙⵏ [winnsn] "le leur").

d. Les mots outils

Les mots outils sont un ensemble de mots amazighes qui ne sont ni des noms, ni des verbes, et jouent un rôle d'indicateurs grammaticaux au sein d'une phrase. Cet ensemble est constitué de plusieurs éléments à savoir: les particules d'aspect, d'orientation et de négation; les adverbes de lieu, de temps, de quantité et de manière; les prépositions; les subordonnants et les conjonctions (Boukhris *et al.*, 2008).

3. Vers un module d'analyse morphologique de la langue amazighe

3.1. La plateforme NooJ

NooJ, publié en 2002 par Max Silberztein (Silberztein, 2007), est un environnement de développement linguistique qui permet de construire et de gérer des dictionnaires électroniques et de grammaires formelles à large couverture afin de formaliser les différents phénomènes linguistiques qui sont : l'orthographe, la morphologie (flexionnelle et dérivationnelle), le lexique (de mots simples, mots composés et expressions figées), la syntaxe (locale, structurelle et transformationnelle), la désambigüisation, la sémantique et les ontologies. Ces descriptions, formalisées à l'aide des machines à états finis tels que les automates à états finis, les transducteurs à états finis et les réseaux de transition récursifs ; peuvent ensuite être appliquées pour traiter des textes et corpus de taille importante afin de localiser les modèles morphologiques, lexicologiques et syntaxiques, lever les ambiguïtés et étiqueter les mots simples et composés.

Le module morphologique de NooJ, utilisé tout au long de cet article, se base sur des opérateurs de transformations (eg. <L> : déplacement vers la gauche, <RW> : aller à la fin du mot, etc) à l'intérieur des formes et des graphes morphologiques décrivant des règles grammaticales à large couverture. A l'aide de ces opérateurs, NooJ permet d'effectuer des recherches et des traitements dans les textes à partir d'expressions régulières et associent leurs reconnaissances à la liste d'annotations correspondante (étiquette grammaticale, des informations sémantiques, la traduction, etc.). Ces transformations fonctionnent sur une pile et

nécessitent un temps de transformation en O(n). Ainsi, elles garantissent une correspondance en un temps linéaire entre le lemme et sa forme flexionnelle ou dérivationnelle. Quant aux grammaires morphologiques, elles sont construites en utilisant l'éditeur de graphes de NooJ et représentées sous forme de transducteurs à états finis. Elles représentent des séquences de lettres et associent leurs reconnaissances à la production des informations lexicales correspondantes.

3.2. Construction de notre lexique amazighe

Bien qu'il soit aujourd'hui clair que le lexique électronique est une composante essentielle des systèmes TAL, les ressources disponibles pour l'amazighe sont très rares et ne couvrant pas les informations morphologiques tels que les formes fléchies, dérivés, etc. Dans ce cadre, nous avons entrepris la construction d'un lexique électronique morphologique qui sera utile pour le développement futur de l'amazighe. Pour y parvenir, nous avons, en premier lieu, développé manuellement une liste de lemmes en exploitant différentes sources d'informations lexicales à savoir : le vocabulaire des médias (Ameur *et al.*, 2009a), le vocabulaire grammaticale (Ameur *et al.*, 2009b), le lexique scolaire (Agnaou *et al.*, 2011), le manuel de conjugaison amazighe (Labdellaoui *et al.*, 2012), le dictionnaire de Taifi (Taifi, 1988), etc. La liste de lemmes, ainsi construite, est divisée en trois lexiques présentés dans le tableau suivant (voir Tableau 1).

Lexiques	Nombre d'entrées
NAmLex (Noun Amazighe Lexicon)	21 371
VAmLex (Verb Amazighe Lexicon)	7 764
PAmLex (Particles Amazighe Lexicon)	1089
Totale	**30 224**

Tableau 1. Le nombre d'entrées pour chaque lexique. (PAmLex[1])

En deuxième lieu, nous avons eu recours au formalisme robuste du dictionnaire inclus dans la plateforme NooJ afin de représenter notre lexique et associé à chacun de ses entrées un ensemble d'informations nécessaires. Cette plateforme nous a permis de représenter les informations lexicales d'une façon complète, efficace et lisible. Ainsi, chaque entrée lexicale présente, généralement, les informations suivantes: le lemme, la catégorie lexicale (N : Nom, V : Verbe, etc.), le type (Com : Commun, Prop : Propre, etc.), le trait syntactico-sémantiques (Anl : Animal, Abs : Abstrait, etc.) et la traduction français et arabe s'ils existent. Et finalement, nous avons terminé par la formalisation de la morphologie (flexionnelle et dérivationnelle) des trois catégories, noms, verbes et pronoms, afin de générer l'ensemble des informations flexionnelles et dérivationnelles pour chaque lemme (Nejme *et al.*, 2013a ; Nejme *et al.*, 2013b ; Nejme *et al.*, 2015 ; Nejme *et al.*, 2016). Par la suite, nous donnons un aperçu de formalisation pour chaque catégorie.

1 Il contient les pronoms et les mots outils.

a. Les noms

Chaque nom, présenté sous forme masculin ou féminin, est associé à :

1. Un modèle de flexion : par modèle de flexion nous désignons l'ensemble des transformations permettant d'obtenir, à partir de chaque entrée nominale, l'ensemble de ses formes fléchis. Ces paradigmes flexionnels incluent le genre (masculin, féminin), le nombre (singulier, pluriel) et l'état (état libre, état d'annexion), ce qui donne, en moyenne, 8 formes fléchies par entrée lexicale.

Exemple :

ⴰⵎⴰⵢⵓ, N + Com+Anl+FLX[2] = PIS: ACaCuC2 + FR = singe + AR = قرد

Parmi les 8 transformations flexionnelles qui sont intégrées dans le paradigme flexionnel "PIS: ACaCuC2", en voici une : « < L > < B > ⴰ < LW > < S > ⵉ < R2 > < B > ⵓ / m + p». Cette transformation NooJ signifie : sauter une lettre vers la gauche (<L>) (abaɣuls), effacer la dernière lettre (<B>) (abaɣls), insérer la voyelle ⴰ [a] (abaɣals), positionner le curseur (|) à la tête du lemme par un déplacement vers la gauche (<LW>) (|abaɣas), effacer la lettre suivante (<S>) (|baɣas), insérer la voyelle ⵉ [i] (i|baɣas), sauter deux lettres vers la droite (<R2>) (iba|ɣas), effacer la dernière lettre (<B>) (ib|ɣas) et enfin insérer la voyelle ⵓ [u] (ibu|ɣas).

Cette opération permet de générer la forme suivante : ⵉⵏⵓⵢⴰⵙ [ibuɣas] "singes" qui sera associée aux informations flexionnelles: N+Com+Anl+m+p, i.e. nom commun (N+Com), trait syntactico-sémantique animal (+Anl) et fléchie en masculin (m) pluriel (p).

2. Un modèle de dérivation : par modèle de dérivation nous désignons l'ensemble des transformations permettant d'obtenir, à partir d'une entrée nominale ou verbale[3], la forme dérivée correspondant. Pour les formes dérivés à base verbale, nous avons associé un ensemble des verbes à des descriptions morphologiques pour représenter l'ensemble des déverbaux (i.e. les noms dérivés à base verbale). Ces noms peuvent être des noms d'actions, des noms d'agents, des noms d'instruments ou des noms de qualités (Nejme *et al.*, 2013a). Par la suite, chacun de ces noms est associé à la flexion correspondante afin de générer l'ensemble des formes fléchis suivant le genre, le nombre et l'état.

2 FLX : fonctionnalité permettant la description des formes fléchies potentielles à partir d'un lemme.
3 Dérivation nominale à base verbale.

Exemple :

ⵜⴰⵎⵎⵉⵔⵜ, N + DRV[4] = tanmmirt:CCVC'C'C_skirry + FR = Merci

L'expression «+DRV=tanmmirt:CCVC'C'C_skirry» désigne un appel au modèle de dérivation «tanmmirt» à utiliser pour produire le verbe ⵙⵏⵎⵎⵔ [snimmr] "remercier". Ce modèle de dérivation fait usage des règles de flexions introduites par ":CCVC'C'C_skirry" pour conjuguer le verbe dérivé généré.

b. Les verbes

Similairement aux noms, chaque verbe, ramené à la 1ème personne du singulier du mode impératif, est associé à :

3. Un modèle de flexion : les paradigmes flexionnels décrits dans ce modèle de flexion incluent l'aspect (aoriste, accompli, etc.), le mode (indicatif, impératif et participiale), les désinences, l'indice de personne, le genre et le nombre, ce qui donne, en moyenne, 130 formes fléchies par entrée lexicale.

Exemple :

ⴽⵛⵔ, V + Simp+Tril+Reg+Intr + FLX = CCC_bgs + FR = entrer + pénétrer + s'introduire+AR = دخل+ ولج

Parmi les 130 transformations flexionnelles qui sont intégrées dans le paradigme flexionnel "CCC_bgs", en voici une : « ⵖ<L2><D> /Inacc+1+m+s+Ind». Cette transformation NooJ signifie : insérer la lettre ⵖ [γ] à la fin du lemme (kcm|γ)[5], sauter deux lettres vers la gauche (<L2>) (kc|mγ) et enfin dupliquer le dernier caractère (<D>) (kcc|mγ).

Cette opération permet de générer la forme suivante : ⴽⵛⵛⵔⵖ [kccmγ] "j'entre" qui sera associée aux informations flexionnelles: V+Simp+Tril+Reg+Intr+Inacc+1+m+s+Ind, i.e. verbe simple (V+Simp), régulier (+Reg), Intransitif (+Intr) de type tritlitère (+Tril), conjugué à la première personne (+1) du masculin (m) singulier (s) à l'inaccompli (+Inacc) de l'indicatif (+Ind).

4. Un modèle de dérivation : ce modèle permet d'obtenir, à partir d'une entrée verbale ou nominale[6], la forme dérivée correspondante.

Exemple : (nous reprenons l'exemple traité ci-dessus pour une dérivation à base verbale)

ⴽⵛⵔ, V+Simp+Tril+Reg+Intr+FLX=CCC_bgs+DRV = Forme1_ss:CC'C'C_sffy+FR = entrer+pénétrer+s'introduire+AR = دخل + ولج

L'expression «+DRV= Forme1_ss:CC'C'C_sffy» désigne un appel au modèle de dérivation «Forme1_ss» à utiliser pour produire la forme factitive ⵙⵙⴽⵛⵔ [sskcm] "faire entrer".

4 DRV : fonctionnalité permettant la description des formes dérivés potentielles à partir d'un lemme.
5 Le curseur (|) est initialement placé à la fin du mot.
6 Dérivation verbale à base nominale.

Ce modèle de dérivation fait usage des règles de flexions introduites par ":CC'C'C_sffy" pour conjuguer la forme dérivé généré suivant l'aspect, le mode, les désinences, l'indice de personne, le genre et le nombre.

c. Les pronoms

En suivant la même démarche utilisée pour les noms et les verbes, nous avons associé pour les entrées pronominales du type démonstratif, possessif ou interrogatif, un mode de flexion permettant de générer les formes fléchis correspondantes.

3.3. Analyse morphologique et définition des règles grammaticales

La langue amazighe est une langue riche en termes de flexions, de dérivations et de complexités qu'elle produit. Son analyse morphologique se déroule en deux phases (voir Figure1).

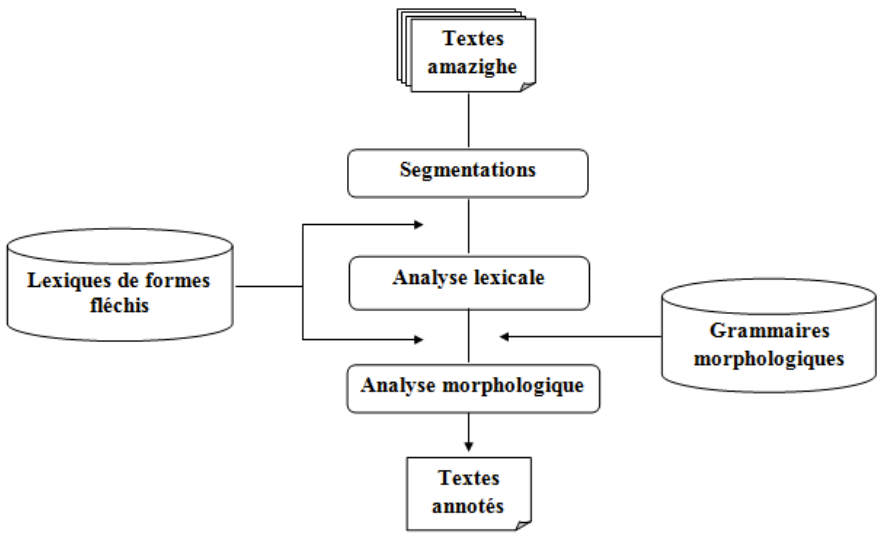


Figure 1. Architecture de l'analyseur morphologique.

Tout d'abord, un système d'analyse linguistique, qui consiste à tester l'appartenance de chaque mot du texte à nos lexiques, est appliqué à l'ensemble de nos textes afin d'associer chaque forme reconnue à l'ensemble d'informations linguistiques potentielles: lemme, étiquette grammaticale, genre et nombre, information syntactico-sémantique, etc.

Ensuite, une phase d'application de règles morphologiques permet de:

(1) Donner tous les découpages potentiels en morphèmes : du fait que notre analyseur sera destiné à des applications qui nécessitent des analyses fines au niveau du mot, il ne s'agira pas seulement de vérifier si le mot appartient aux lexiques, mais aussi de donner tous les découpages potentiels en morphèmes et d'associer pour chaque découpage retenu, l'ensemble d'information morpho-syntaxiques correspondantes ;

(2) Traiter un ensemble de contraintes morphologiques, orthographiques et phonologiques: malgré tous les efforts pour améliorer nos lexiques, l'analyse automatique est régulièrement confrontée aux problèmes liés à un ensemble de contraintes morphologiques, orthographiques ou phonologiques. Dans le but d'améliorer la robustesse de notre analyseur, nous avons construits un ensemble de grammaires permettant de traiter ces contraintes. Pour illustrer ce propos, nous présentons, ci-dessous, un exemple de grammaire permettant de reconnaître, de découper et d'annoter les noms de parentés qui viennent attachés avec les pronoms possessifs qui les déterminent (voir Figure 2).

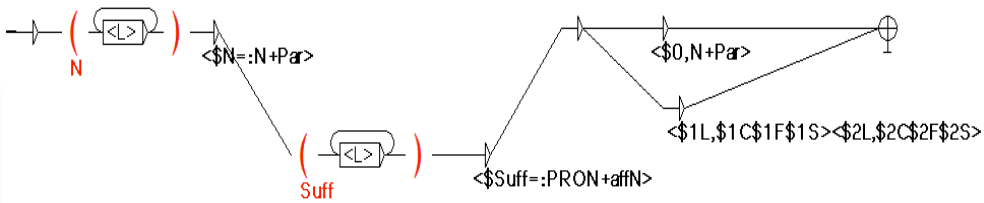


Figure 2. Exemple de grammaire de tokenisation des noms de parenté.

Le graphe simplifié ci-dessus montre que l'acceptation d'une entrée formée par l'agglutination d'une suite de lettres (<L>*) sauvegardée dans la variable ($N), suivie par une deuxième suite de lette sauvegardée dans la variable ($Suff) est liée à la vérification des deux contraintes lexicales suivantes:

- d'une part la variable '$N' doit représenter un nom de parenté dans notre lexique (<$N=:N+Par>) ;
- d'autre part la deuxième variable '$Suff' doit être valider en tant que pronom affixe (<$Suff=:PRO+affN>).

En cas de validité des deux contraintes lexicales citées, nous produisons l'ensemble des informations morpho-syntaxiques rattachées à chaque morphème de la segmentation retenue.

4. Conclusion et perspectives

L'objectif principal visé par ce travail était de créer un module d'analyse lexicale et morphologique pour la langue amazighe. Cet objectif est accompli par une formalisation des quatre catégories morpho-syntaxiques de l'amazighe accompagné d'une chaîne d'analyse morphologique. Le module ainsi construit profite des apports des modèles à états finis au sein de l'environnement linguistique de développement NooJ en faisant appel à des règles grammaticales à large couverture.

L'évaluation de la couverture lexicale de ce module est entreprise en effectuant l'analyse d'un corpus de comptes pour enfants. Ce corpus comporte environ 13 876 formes différentes. Le résultat de l'analyse montre que le vocabulaire du corpus est reconnu à environ 98 % par nos ressources lexicales et morphologiques. L'ensemble des formes non reconnues contient

principalement des entités nommés (ⵎⵉⵚⵔ [misr] “Egypt”), des emprunts (ⴱⵣⵣⴰⴼ [bzzaf] “a lot”) et des nouveaux mots qui n’existent pas dans nos lexiques.

Références

Ameur M., Bouhjar A., Boukhris F., Boukouss A., Boumalk A., Elmedlaoui M., Iazzi E. (2006). *Graphie et orthographe de l’amazighe*. Rabat, Maroc : IRCAM.

Ameur M., Bouhjar A., Boumalk A., El Azrak N., Laabdelaoui R. (2009a). *Vocabulaire des médias (Français-Amazighe-Anglais-Arabe).* Série : Lexiques N°3, IRCAM, Rabat, Maroc.

Ameur M., Bouhjar A., Boumalk A., El Azrak N., Laabdelaoui R. (2009b). *Vocabulaire grammatical*. Série : Lexiques N°5, IRCAM, Rabat, Maroc.

Agnaou, F., Bouzandag, A., El Baghdadi, M., El Gholb, H., Khalafi, A., Ouqua, K., Sghir, M. (2011). *Lexique scolaire*. IRCAM, Rabat.

Ataa Allah F., Boulaknadel S. (2010). *Online Amazigh Concordancer*. Proceedings of International Symposium on Image Video Communications and Mobile Networks. Rabat, Maroc.

Boukhris F., Boumalk A., Elmoujahid E., Souifi H. (2008). *La nouvelle grammaire de l’amazighe*. Rabat, Maroc: IRCAM.

Boukous A. (1995), Société, langues et cultures au Maroc: Enjeux symboliques, Casablanca, Najah El Jadida.

Boulaknadel, S., Ataa Allah, F. (2011). *Building a standard Amazigh corpus*. Proceedings of the International Conference on Intelligent Human Computer Interaction, Prague, Tchec.

Es Saady Y., Rachidi, A., El Yassa, M., Mammas, D. (2010). *Printed Amazighe Character Recognition by a Syntactic Approach Using Finite Automata*. International Journal on Graphics, Vision and Image Processing, vol. 10 no. 2, pp. 1–8.

Fakir M., Bouikhalene B., Moro K. (2009). *Skeletonization methods evaluation for the recognition of printed tifinaghe characters*. In Proceedings of the 1er Symposium International sur le Traitement Automatique de la Culture Amazighe, pp. 33—47. Agadir, Morocco.

Greenberg J. (1966). *The Languages of Africa*. The Hague.

Kamel S. (2006). *Lexique Amazighe de géologie*. Rabat, Maroc: IRCAM.

Laabdelaoui R., Boumalk A., Iazzi M., Souifi H. and Ansar K. (2012). *Manuel de conjugaison de l’Amazighe*. IRCAM, Rabat, Morocco.

Max S. (2007). *An Alternative Approach to Tagging*. NLDB 2007: 1-11.

Nejme F., Boulaknadel S., Aboutajdine D. (2013a). *Analyse Automatique de la Morphologie Nominale Amazighe*. Actes de la conférence du Traitement Automatique du Langage Naturel (TALN). Les Sables d’Olonne, France.

Nejme F., Boulaknadel S., Aboutajdine D. (2013b). *Finite State Morphology for Amazighe Language*. In Proceeding of International Conference on Intelligent Text Processing and Computational linguistics (CICLing). Samos, Greece.

Nejme F., Boulaknadel S., Aboutajdine D. (2012). *Vers un dictionnaire électronique de l'Amazighe*. Actes de la Conférence Internationale sur les Technologies d'Information et de Communication pour l'AMazighe (TICAM). Rabat, Maroc.

Nejme F., Boulaknadel S., Aboutajdine D. (2015). *Automatic Processing of Amazighe Verbal Morphology*: a New Approach for Analysis and Formalization. In International Review on Computers and Software (IRECOS).

Nejme F., Boulaknadel S., Aboutajdine D. (2016). *AmAMorph: Finite State Morphological Analyzer* for Amazighe. CIT. Journal of Computing and Information Technology, vol. 24, no. 1, pp. 91-110.

Ouakrim O. (1995). *Fonética y fonología del Bereber*. Survey at the University of Autònoma de Barcelona.

Outahajala M., Zenkouar L., Rosso P., Martí M. (2010). *Tagging Amazigh with AncoraPipe*. Actes de Semitic Languages Workshop, 7th International Conference on Language Resources and Evaluation. pp. 52-56.

Taghbalout I., Ataa Allah F., El Marraki M. (2015). *Amazigh Noun Inflection in the Universal Networking Language*. International journal of education and information technologies.

Taifi M. (1988). *Le lexique berbère (parlers du Maroc central).*

Talha M., Boulaknadel S., Aboutajdine D. (2015). *L'apport d'une approche symbolique pour le repérage des entités nommées en langue amazighe*. Acte de la conférence internationale sur l'extraction et la gestion des connaissances (EGC).

Etiquetage morphosyntaxique de l'amazighe: particularités et enjeux

Samir Amri[1] Lahbib Zenkouar[1] Mohamed Outahajala[2]
amri.samir@gmail.com
[1] Ecole Mohammadia d'Ingénieurs, Rabat, Maroc
[2] Institut Royal de la Culture Amazighe (IRCAM), Rabat, Maroc

Résumé

L'objectif principal de ce papier est de présenter les particularités et les enjeux relatifs à la tâche d'étiquetage morphosyntaxique. En effet, l'étiquetage morphosyntaxique est une partie vitale de toute application du Traitement Automatique des Langues Naturelles (analyseur syntaxique, traducteur automatique, correcteur orthographique…), car la performance de toute application dépend, entre autres, de la performance de l'étiqueteur morphosyntaxique qu'elle utilise. Ainsi, et afin de réaliser un étiqueteur morphosyntaxique efficace, on doit s'intéresser à améliorer la qualité des trois phases suivantes: la phase de segmentation, la phase d'organisation des unités lexicales et la phase de désambiguïsation.

1. Introduction

L'amazighe est parmi les langues peu dotées et les moins utilisées sur Internet, d'où la motivation et la nécessité de son informatisation et de son développement en Traitement Automatique des Langues Naturelles (TALN). D'ailleurs beaucoup de recherches ont dirigé cette tâche du TALN et ont abouti à diverses approches et algorithmes qui ont conduit fréquemment à des applications et à des systèmes sophistiqués.

D'un point de vue général, pour la mise en œuvre d''outils du TALN, les chercheurs ont besoin:

- d'unités de base pour la segmentation des phrases et des mots et de l'analyse morphologique, syntaxique ou sémantique.
- des ressources linguistiques (dictionnaires et des agrégats, données lexicales, corpus...).
- des expertises au niveau linguistique ou au niveau d'apprentissage automatique (Machine Learning en Anglais).

Au niveau de cet article on va se focaliser sur la discipline de l'étiquetage morphosyntaxique qui est une étape indispensable et primordiale pour la réalisation de la plupart des applications du TALN, car il peut déterminer la catégorie grammaticale des mots de texte et la description des différentes unités de base dans les applications grand public telles que l'analyse syntaxique, la génération automatique des résumés et la recherche d'information...

etc. Il est également très utile dans le traitement des mots pour les systèmes d'optimisation des performances et la reconnaissance vocale. En général, l'étiquetage morphosyntaxique est une étape nécessaire et difficile à faire. Ainsi, nous avons décidé de mettre l'accent sur ce problème en particulier pour la langue amazighe.

La suite de cet article est structurée comme suit : la deuxième section est consacrée à un état de l'art sur la langue amazighe et son traitement automatique. La troisième section est focalisée l'étiquetage morphosyntaxique. La quatrième section est dédiée à une discussion sur les prérequis nécessaires et les points à améliorer pour avoir des systèmes d'étiquetage morphosyntaxiques complets robustes et surtout efficaces pour la langue amazighe. Enfin on conclura avec un ensemble de perspectives pour les travaux futurs dans le domaine du traitement automatique de la langue amazighe.

2. Amazighe et TALN

2.1 Aperçu sur la langue amazighe

L'amazighe est parlé sous forme de plusieurs dialectes et parlers. Ces derniers sont utilisés sur un grand territoire qui couvre de nombreux pays: Egypte, Libye, Tunisie, Algérie, Maroc, Mali, Niger, Mauritanie. Cependant l'Algérie et le Maroc sont les deux pays où est concentré le plus grand nombre d'Imazighen dans le sens qu'être amazighe c'est parler un des parlers de la langue.

Selon les régions, ces parlers prennent des noms différents. Ainsi en Algérie, nous nous retrouvons notamment avec les parlers Kabyle, Mozabite et Chaoui. Au Maroc, il y a trois parlers principaux:

Tarifit au nord du Maroc, Tachelhit au sud-ouest du royaume et Tamazight au Maroc central.

Malgré des nombreuses recherches, la langue amazighe est considérée comme une langue difficile à maîtriser à cause de sa richesse morphologique. Les travaux de recherche dans le TALN ont abordé des problématiques variées comme la morphologie, la traduction automatique, l'indexation des documents, etc.

Au cours de ce passage nous présenterons les particularités de la langue amazighe ainsi que certaines de ses propriétés morphologiques et syntaxiques.

La création de l'Institut Royal de la Culture Amazighe(IRCAM) en 2001 et l'officialisation de la langue amazighe en 2011 ont permis la promotion de la langue amazighe. Ainsi, elle a obtenu une orthographe officielle (Ameur *et al.*, 2004), un codage approprié dans le standard Unicode (Andries, 2008; Zenkouar, 2008) et des structures linguistiques (Ameur *et al.*,2004; Boukhris *et al.*, 2008).

La langue amazighe possède sa propre graphie qui est le Tifinaghe, un système alphabétique standard plus adéquat et utilisable pour tous les parlers amazighes actuels. Ainsi en 2003, l'IRCAM a développé un système d'alphabet sous le nom de Tifinaghe-IRCAM. L'alphabet standardisé par l'IRCAM est basé sur un système graphique à tendance phonologique, cet

alphabet comporte :

- 27 consonnes dont : les labiales (ⵀ, ⵃ, ⵆ), les dentales (ⵜ, ⵟ, ⵠ, ⵡ, ⵍ, ⵏ, ⵇ, ⵅ), les alvéolaires (ⵚ, ⵣ, ⵛ, ⵥ), les palatales (ⵞ, ⵊ), les vélaires (ⴽ, ⵅ), les labiovélaires (ⴽⵯ, ⵅⵯ), les uvulaires (ⵖ, ⵅ, ⵄ), les pharyngales (ⵄ, ⵃ) et la laryngale (ⴼ).
- 2 semi-consonnes : ⵢ et ⵡ.
- 4 voyelles : trois voyelles pleines ⴰ, ⵉ, ⵓ et la voyelle neutre ⵔ qui a un statut assez particulier en phonologie amazighe.

D'ailleurs c'est la translitération en alphabet latin qui est utilisée dans tous les exemples présentés dans cet article.

Par ailleurs, dans le lexique de la langue amazighe, on distingue trois catégories principales de mots :

Les verbes, les noms et les particules (Boukhris *et al.*, 2008) qui se subdivisent elles-mêmes en différentes sous catégories: préposition, conjonction, pronom, article, interjection et adverbe :

- Le nom est soit au masculin, soit au féminin. Il est au pluriel ou au singulier: le pluriel commence à partir de deux comme en Français. Le nom est soit à l'état libre ou à l'état d'annexion.
- Par exemple pour le nom masculin : afus /ifassn (main/mains), igr/igran (champ/champs), pour le nom féminin: tuzzalt/tuzzalin (couteau/couteaux), tasarut/tisura (clef/clés).
- Le verbe se construit généralement par l'affixation et la composition. Certains verbes sont des dérivations par affixation (préfixes, suffixes), d'autre verbes ne sont pas nécessairement dérivés de noms, ils sont composés soit à partir d'un verbe et d'un nom, soit à partir de deux verbes, sans oublier bien évidemment les aspects de la conjugaison qui impactent parfois la morphologie du verbe d'une façon significative.
- Exemple du verbe en amazighe : sw(boire), ddu (aller), rwl(courir).
- Les pronoms sont isolés des mots auxquels ils se réfèrent. Les pronoms en langue amazighe sont soit démonstratifs, exclamatifs, indéfinis, interrogatifs, personnels, possessifs ou relatifs.
- Les adverbes sont subdivisés en adverbes de lieu, de temps, de quantité, de manière et les adverbes interrogatifs.
- Les prépositions sont un ensemble de caractères indépendants par rapport au nom qu'elles précèdent; cependant si la préposition est suivie d'un pronom personnel, la préposition et le pronom personnel forment une seule chaîne délimitée par des blancs ou bien un blanc et une marque de ponctuation.
- Les particules sont toujours isolées, elles sont de plusieurs types:

    - Les particules aspectuelles telles que «ar, ad».

- La particule de négation telle que « ur ».
- Les particules d'orientation telle que « s ».
- La particule de prédication telle que « d».

- Les déterminants prennent toujours la forme d'un seul mot délimité par deux espaces, ils sont divisés en articles, démonstratifs, exclamatifs, articles indéfinis, interrogatifs, chiffres ordinaux, possessifs, présentatifs et quantificateur.
- Les marques de ponctuation en amazighe marocain sont similaires aux marques de ponctuation adoptées par les langues internationales, elles ont les mêmes fonctions.

2.2 Traitement Automatique de la Langue Amazighe (TALAM)

Le Traitement Automatique d'une Langue Naturelle (TALN) est divisé d'une façon générale en deux parties :

- Traitement de langue: concerne les systèmes capables de se comporter comme des lecteurs/auditeurs.
- Génération de langue : concerne les systèmes capables de se comporter comme des rédacteurs/producteurs.

Après cette subdivision, on entrevoit des niveaux dans le TALN :

- Le niveau phonologie: interpréte le discours à travers les mots.
- Le niveau morphologique: traite la composition des mots (préfixe, suffixe, radical, ...).
- Le niveau lexical: donne un sens au mot pris individuellement.
- Le niveau syntaxique: découvre la structure grammaticale de la phrase.
- Le niveau sémantique: traite le sens des mots et des phrases.
- Le niveau conversation: traite le sens global des corpus. Il ne considère pas un texte comme une concaténation de phrases, mais comme un ensemble pourvu de sens.
- Le niveau pragmatique: explicite les sens implicites des phrases et mots.

En ce qui concerne le TALAM, la langue amazighe ne possède pas suffisamment des ressources linguistiques et d'outils TALN (Outahajala *et al.* 2015). Toutefois on va lister quelques travaux déjà faits pour le TALAM :

- Le développement d'outils adaptés au traitement de l'amazighe (Rachidi et Mammass, 2005).
- La création des claviers et polices de caractères dédiés à l'écriture Tifinaghe (IRCAM, 2003b; IRCAM, 2004).
- La réalisation des travaux de translittération des textes écrits en alphabet tifinaghe vers l'alphabet arabe ou latin (Ataa Allah *et al.* 2013).
- La construction d'un grand corpus annoté pour la langue amazighe (Outahajala *et al.*, 2014).
- L'élaboration de système de reconnaissance des caractères Tifinaghes (Ait Ouguengay *et al.*, 2009).

- L'analyseur morphologique pour les noms amazighes (Raiss & Cavalli Sforza, 2012).
- Le conjugueur des verbes de la langue amazighe. (Ataa Allah et Boulaknadel, 2014).
- Le pseudo-racineur (Ataa Allah et Boulaknadel, 2010).
- Le concordancier (Boulaknadel, 2009), permettant la recherche d'un mot quelconque dans un ensemble de textes afin d'étudier son emploi.

De ce qui précède on peut constater que le domaine du TALAM a besoin de vision et de stratégie de tout le monde (chercheurs, linguistes…) pour réussir ce grand chantier et d'apporter à la communauté scientifique et au grand public des systèmes et des projets pertinents et de grande valeur ajoutée.

3. L'étiquetage morphosyntaxique de la langue amazighe

Il s'agit d'un processus de détecter la catégorie morphosyntaxique d'un mot dans un contexte, cette action est non triviale du traitement automatique de la langue écrite. En effet rendre un ordinateur capable de connaître la catégorie grammaticale d'un mot exige de mettre en œuvre des méthodes sophistiquées, en particulier pour les mots ambigus, c'est-à-dire susceptibles d'appartenir à plusieurs catégories différentes. Les systèmes automatiques dédiés à cette activité sont appelés des étiqueteurs morphosyntaxiques (Part-Of-Speech tagger en Anglais).

Ceux-ci consistent à affecter des étiquettes morphosyntaxiques propres à chaque mot d'une phrase d'un texte (catégorie grammaticale, informations morphologiques comme le genre, le nombre, l'état...etc). L'étiquetage correct par exemple de la phrase (*idda yidir s tmzgida*) est comme suit : *idda.Verbe yidir*. *Nom propre s.préposition tmzgida.Nom*. La principale difficulté de l'étiquetage morphosyntaxique vient du fait que les mots de la langue sont ambigus, c'est à dire que l'on peut affecter plusieurs étiquettes à un mot donné de la phrase.

Un étiqueteur morphosyntaxique doit donc effectuer une phase de désambigüisation afin de sélectionner une séquence d'étiquettes possibles pour la séquence de mots de la phrase, et si possible la séquence correcte.

D'ailleurs l'étiquetage morphosyntaxique a été largement étudié par le passé, il est maintenant considéré comme un problème relativement résolu pour quelques langues comme l'Anglais et le Français. Les performances des étiqueteurs actuels de ces langues étant très élevées (environ 97,50% de mots correctement étiquetés).

Pour aborder cette discipline, plusieurs approches ont été proposées pour annoter automatiquement les mots d'un texte (figure1).

Le mécanisme de l'étiquetage morphosyntaxique se base généralement sur l'hypothèse que la catégorie d'un mot dépend de son contexte local, qui peut par exemple se réduire au mot ou aux deux qui le précèdent.

Dans ce qui suit nous allons présenter différentes méthodes d'étiquetage morphosyntaxique, et effectuer un bref recensement des étiqueteurs qui existent en particulier pour la langue amazighe.

Il existe deux grandes familles d'étiqueteurs :

- Les étiqueteurs symboliques sont ceux qui appliquent des règles qui leur ont été communiquées par des experts humains. Dans ce type d'étiqueteurs, il y a très peu d'automatisation; c'est le designer qui manipule toutes les règles d'étiquetage et qui fournit au besoin une liste des morphèmes. La conception n'est pas automatisée : l'étiqueteur fournit un étiquetage automatique une fois ses règles élaborées. La conception d'un tel étiqueteur est longue et coûteuse. De plus, les étiqueteurs ainsi conçus ne sont pas facilement portables, c'est-à-dire ils ne sont efficaces que pour une langue donnée et un domaine donnée (exemple: la finance, la politique, etc.).
- Les étiqueteurs avec apprentissage automatique (Machine Learning en Anglais) sur lesquels nous allons nous concentrer dans la suite de cette étude. Parmi les étiqueteurs de ce type, il existe deux grands types: les étiqueteurs supervisés qui apprennent à partir de corpus pré-étiquetés (Brill, 1993 ; Khoja, 2001 ; Diab et al., 2004) et les étiqueteurs non supervisés qui apprennent à partir de corpus bruts sans information additionnelle. Qu'ils soient supervisés ou non, les étiqueteurs avec apprentissage peuvent être regroupés en trois familles: systèmes à base de règles, statistiques ou neuronal.

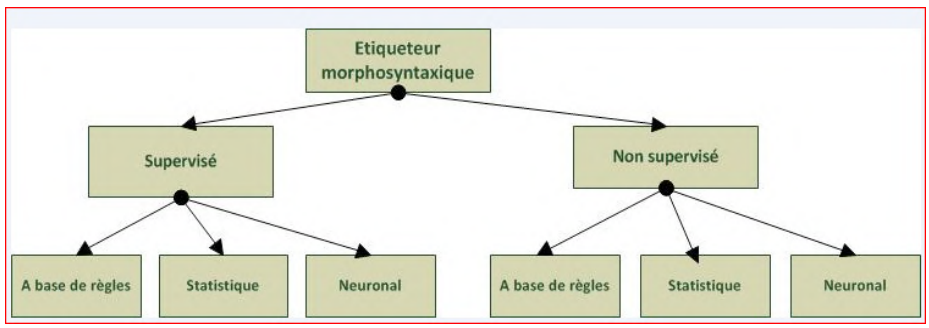


Figure1 : Les différentes méthodes d'étiquetage morphosyntaxique avec apprentissage automatique

L'étiquetage à base des règles possède les points forts suivants: son fondement linguistique, ses règles lisibles et modifiables manuellement, sa facilité à la compréhension des erreurs, sa base de connaissances qui peut être modifiée: suppression de règles ou ajout de nouvelles règles. La grande faiblesse de cet étiquetage réside dans le problème de contradiction entre les règles, ce qui nécessite de disposer des dictionnaires de règles qui est un travail manuel et couteux. D'un point de vue général ce sont des systèmes plus rapides mais moins précis.

Alors que l'intérêt de l'approche statistique de l'étiquetage réside dans le fait qu'on peut déterminer correctement la catégorie d'un mot sans l'avoir jamais vu auparavant.

4. Étiqueteurs morphosyntaxiques

Au début de ce paragraphe on va lister quelques étiqueteurs morphosyntaxiques qui sont disponibles pour la recherche scientifique (tableau 1), et qui jouissent d'un grand avantage du fait qu'ils sont tous indépendants de la langue, il suffit pour les implémenter d'avoir un corpus pour l'apprentissage et un autre pour les tests et un lexique pour quelques-uns (TreeTagger).

Etiqueteurs	Référence	Technique utilisée
TreeTagger (supervisé)	(Schmidt , 1994)	Modèle de Markov Caché (MMC) et Arbres de décisions
Trigrams'n'Tags(TnT) (supervisé)	(Brants , 2000)	Modèle de Markov Caché (MMC)
SVMTool (supervisé)	(Giménez and Marquez , 2006)	Séparateurs à Vaste Marge (SVM)
CRF++ (supervisé)	(Lafferty, J. McCallum, A. and Pereira, F. 2001)	Champs Markoviens Conditionnels (CRF)
Yamcha (supervisé)	TakuKudo, Yuji Matsumoto (2000)	Séparateurs à Vaste Marge (SVM)
MXPOST (supervisé)	(Ratnaparkhi, 1994)	Entropie maximale
Stanford Pos Tagger (supervisé)	(Toutanova and Manning, 2000)	Entropie maximale
Unsupos (non supervisé)	(Chris Biemann's, 2007)	Viterbi
Brill (supervisé)	(Brill, 1992)	Règles lexicales + Règles contextuelles

Tableau 1 : Quelques étiqueteurs disponibles à la recherche avec référence et technique d'apprentissage automatique

Les mots inconnus, semblent être un problème pour tous les étiqueteurs basés sur des algorithmes d'apprentissage qui produisent des modèles de langage. Cependant certains mentionnés peuvent être modifiés pour tenir compte également des connaissances lexicales et effectuer la lemmatisation également, en particulier Brill et CRF++. Stanford et MXPOST peuvent être extensibles aussi bien, mais leur code est plutôt complexe, ce qui rend probablement le développement d'extensions difficile. Quant à Unsupos, l'approche de l'apprentissage non supervisé reste une piste si le corpus annoté n'est pas disponible pour la langue à étudier.

En terme des performances, les modèles probabilistes discriminants comme les modèles de maximum d'entropie (Ratnaparkhi, 1994; Toutanova *et al.*, 2003), les séparateurs à vaste marge (Giménez et Márquez., 2004) ou les champs markoviens conditionnels (Tsuruoka *et al.*, 2009) fournissent de bons résultats en étiquetage morphosyntaxique.

4.1 Corpus de travail et jeu d'étiquettes

Un corpus est une collection de divers matériaux rassemblés selon un ensemble de critères afin qu'il soit représentatif et balancé.

L'utilisation des corpus constitue une phase critique des systèmes du TALN basés sur des méthodes statistiques (Habash et Rambow, 2005).

Les corpus les plus populaires pour l'anglais sont le Brown Corpus (Kurcera et Francis, 1967) qui contient environ un million de mots et le Penn Treebank qui est un corpus commercialisé par le Consortium des Données Linguistiques (LDC).

Pour la langue arabe, le premier corpus annoté réalisé est celui de Khoja et ses co-auteurs, ce corpus contient 50000 mots annotés (Khoja *et al.*, 2001). D'autres corpus sont utilisés tels le Penn Arabic Treebank (Maamouri *et al.*, 2004) et le Prague Arabic Dependency Treebank (Smrz et Hajic, 2006).

Pour les langues disposant de peu de ressources électroniques et peu informatisées comme la langue amazighe, la motivation principale d'avoir un corpus annoté est obtenir des données d'entraînement pour les étiqueteurs morphosyntaxiques d'une part et d'autre part fournir aux applications du TALAM un outil de base.

Malgré les différentes recherches effectuées sur le traitement automatique de la langue amazighe, il est difficile de trouver des ressources linguistiques toutes faites, on peut citer le corpus annoté manuellement (Outahajala *et al.*, 2015). Ce corpus contient 20k mots utilisant un jeu d'étiquette (Tagset en Anglais) décrit dans le tableau 2 , il s'agit d'une étape importante pour un travail d'étiquetage lexical qui doit être basé sur les classes de mots de la langue et doit refléter toutes les relations morphosyntaxiques des mots du corpus amazighe.

Dans le cas de la langue amazighe, la question de la classification des catégories grammaticales est une tâche difficile et toujours en débat au sein de l'IRCAM. Dans ce sens, un colloque sera organisé par l'IRCAM en Décembre prochain pour traiter de la question des catégories grammaticales amazighes. Le jeu d'étiquette doit représenter la richesse des informations lexicales, ainsi que l'information nécessaire à la désambiguïsation.

Etiquette	attributs et sous attributs avec le nombre des valeurs
Nom	genre (3), nombre (3), état (2), dérivation (2), POS sous classification (4), nombre du possesseur (3), genre du possesseur (3), personne (3)
Verbe	genre (3), nombre(3), personne(3), aspect(3), négation(2), forme(2), dérivation(2), voix(2)
Adjectif	genre(3), nombre(3), état(2), dérivation(2), POS sous classification(3)
Pronom	genre(3), nombre(3), personne(3), POS sous classification(7), déictique(3)
Déterminant	genre(3), nombre (3), POS sous classification(11), déictique(3)
Adverbe	POS sous classification(6)
Préposition	genre(3), nombre(3), personne(3), nombre du possesseur(3),genre du possesseur(3)
Conjonction	POS sous classification(2)
Interjection	Focalisateur
Particule	POS sous classification(7)
Focaliseur	Focaliseur
Résiduel	POS sous classification(5), genre(3), nombre(3)
Ponctuation	type de la marque de ponctuation(16)

Tableau 2 : Jeu d'étiquette de base utilisé lors de l'étiquetage morphosyntaxique de l'amazighe

4.2 Étiquetage morphosyntaxique automatique

L'étiquetage morphosyntaxique automatique de la langue est un processus qui s'effectue généralement en 3 étapes :

- La segmentation du texte en unités lexicales.
- L'étiquetage qui consiste à attribuer pour chaque unité lexicale l'ensemble des étiquettes morphosyntaxiques possibles.
- La désambiguïsation qui permet d'attribuer, pour chacune des unités lexicales en fonction de son contexte, l'étiquette morphosyntaxique la plus probable.

4.2.1 Segmentation des unités lexicales

L'étiquetage morphosyntaxique pour l'amazighe reste toujours un sujet d'intérêt pour de nombreux chercheurs du fait de son rôle de brique de base dans de nombreuses applications

du TALN. Bien que de nombreux systèmes aient été réalisés selon des méthodes différentes, les pistes d'amélioration sont encore très ouvertes. Avant d'aborder l'étiquetage morphosyntaxique, il faut préalablement effectuer un prétraitement du texte en entrée : le texte doit être tokenisé, c'est-à-dire segmenté au niveau lexical.

La segmentation est un processus nécessaire dans le traitement morphologique de la langue. Le but de la segmentation est de diviser un texte en une suite de morphèmes afin de préparer le traitement morphosyntaxique (étiquetage ou POS tagging en anglais).

4.2.2 Etiquetage morphosyntaxique pour l'amazighe

Dans ce contexte, (Outahajala *et al.* , 2015) ont conçu et développé deux modèles de classification de séquences pour la langue amazighe, à savoir: les séparateurs à vaste marge (Support Vector Machines, SVMs) en utilisant l'outil open source Yamcha, les champs markoviens conditionnels (Conditional Random Fields, CRFs) en utilisant l'outil open source CRF++ après une phase de segmentation.

Ces modèles utilisés se basent sur la programmation dynamique pour le choix optimal de l'étiquette, et ce en utilisant les propriétés de contexte pour choisir la séquence d'étiquettes maximisant dynamiquement les étiquettes données. Dans leurs expérimentations, ils ont utilisé la technique de 10 fois validation croisée pour évaluer la démarche suivie. Sachant qu'ils ont utilisé un corpus d'environ ~ 20k mots, les résultats obtenus sont approximatifs à l'état d'art des étiqueteurs morphosyntaxiques. Pour améliorer la précision de leurs étiqueteurs morphosyntaxiques une ressource lexicale enrichie avec les étiquettes grammaticales d'environ 8k mots a été utilisée ce qui a permis d'obtenir une performance de 93.82%, soit un gain en précision de 2.64%.

Les Figures 1 et 2 illustrent ces 2 phases indispensables pour la conception et le développement d'un étiqueteur morphosyntaxique :

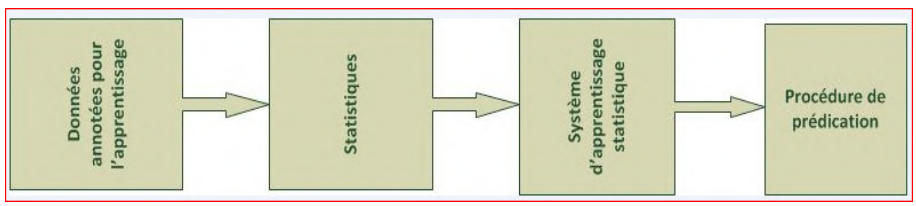


Figure1 : Partie d'apprentissage des étiqueteurs morphosyntaxiques

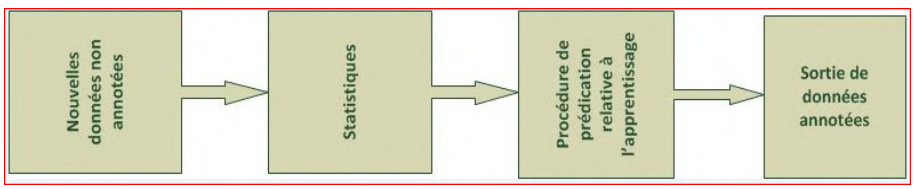


Figure2 : Partie étiquetage des étiqueteurs morphosyntaxiques

Il a par ailleurs été montré que le couplage des modèles d'apprentissage automatique avec des lexiques externes augmente encore la qualité de l'annotation, comme l'illustre (Outahajala *et al.* 2015) pour les CRFs.

4.2.3 La désambiguïsation

Deux problèmes majeurs empêchent les étiqueteurs morphosyntaxiques d'atteindre la précision de 100% : l'ambiguïté des mots et les mots inconnus (Martinez, 2011). Par exemple, le mot amazighe « tazla »peut être aussi bien un nom qu'un verbe; cela dépend du contexte d'utilisation.

Les systèmes d'étiquetage implémentent des algorithmes pour régler la question, ces algorithmes ne sont pas toujours efficaces. Quelques fois, des connaissances sémantiques sont indispensables pour lever des ambiguïtés. Or en étiquetage morphosyntaxique l'accent est mis sur la forme des mots et jamais sur la sémantique qui est un domaine à part.

Les mots inconnus (ou mots hors vocabulaire) sont ceux qui ne se trouvent pas dans le corpus d'apprentissage du système et que le système est censé retrouver.

Pour être robuste face à ces problèmes, la plupart des étiqueteurs utilisent des informations statistiques. Selon (Manning et Schütze, 2000), il y a deux sources possibles d'information pour l'étiquetage : (i) regarder les catégories des mots environnants (ii) regarder la probabilité d'occurrence d'une catégorie lexicale. On peut calculer les probabilités des étiquettes qui correspondent au mot courant, en considérant deux(02) (bigrammes) ou trois (03) (trigrammes) catégories et/ou valeurs de mots situées avant et/ou après. Les trigrammes sont plus efficaces car ils tiennent compte davantage du contexte.

5. Discussion et projets futurs

Les pistes d'amélioration de la segmentation et l'étiquetage morphosyntaxique sont envisageables tout en exploitant l'apprentissage profond (Deep Learning en Anglais). Cet apprentissage est basé sur un ensemble d'algorithmes visant la modélisation des abstractions de haut niveau au sein des données, en utilisant des architectures de modèles composés de multiples transformations non linéaires.

D'un point de vue général, le traitement automatique de la langue amazighe - et en particulier l'étiquetage morphosyntaxique - reste un domaine très ouvert et présente des marges de progression importantes, du fait de la richesse morphologique de cette langue.

Comme nous l'avons mentionné, la segmentation lexicale, une des opérations de base souvent considérée comme triviale dans des langues comme l'anglais ou le français, reste un des problèmes clés de l'amazighe, où de grandes améliorations peuvent encore être apportées.

Dans le futur, on compte travailler sur les axes suivants:

- Construction d'un corpus d'apprentissage amazighe: l'étiquetage supervisé doit être essentiellement accompagné d'un corpus d'apprentissage annoté de qualité. De coup pour obtenir des bons résultats sur n'importe quel texte du test, notre corpus doit être très équilibré et balancé: accueillir des phrases des domaines divers (religion, art, éducation, littérature…).

- Développement d'un étiqueteur amazighe de grande performance tout en exploitant les techniques d'apprentissage automatique. Dans ce projet on va essayer d'exploiter l'approche de combinaison des différentes techniques probabilistes existantes ou nouvelles pour construire un étiqueteur morphosyntaxique de grande précision. On va montrer que cette approche qu'on va appliquer pour la langue amazighe peut être utilisée pour les langues peu dotées.
- Mise en oeuvre d'un système de lemmatisation : le lexique amazighe doit contenir les vrais lemmes des mots pour que le système analyse correctement une phase.
- Mise en oeuvre d'un système de détection des entités nommées (noms propres, villes, offices et institutions…) à partir des textes amazighes.

Conclusion

L'étiquetage morphosyntaxique est la première brique dans la majorité des applications du TALN, la précision de toute application de TALN dépend de la précision de l'étiqueteur.

D'ailleurs différentes approches peuvent être utilisées par les chercheurs pour le développement des étiqueteurs de la langue amazighe.

Dans cet article, on a constaté que la langue amazighe est une langue morphologiquement riche, d'où la nécessité de développement d'un analyseur morphologique tout en exploitant des techniques d'apprentissage automatiques pour la construction des étiqueteurs qui ont des précisions similaires aux étiqueteurs des langues européennes (Anglais, Français, Allemand, …).

En plus, un travail limité a été fait sur la langue amazighe pour la partie de l'étiquetage morphosyntaxique, par conséquent différentes approches peuvent être utilisées pour le développement d'un étiqueteur amazighe robuste et efficace.

Références

Ameur M., Bouhjar A., Boukhris F., Boukouss A. and Boumalk A. (2004) : Initiation à la langue Amazighe. Publications de l'IRCAM.

Ataa Allah F. and Boulaknadel S. (2010). Pseudo-racinisation de la langue Amazighe. In Proceedings of TALN 2010, Montréal, pp.19--23.

Boulaknadel S. (2009). Amazigh ConCorde. An Appropriate Concordance for Amazigh. In Proceedings of 1er Symposium International sur le Traitement Automatique de la Culture AMazighe (SITACAM). Agadir, Morocco.

Boukhris F., Boumalk A., El moujahid E., Souifi. (2008). La nouvelle grammaire de l'Amazighe. Publications de l'IRCAM H.2008.

Giménez J. and Màrquez L. (2004).SVMTool. A General POS Tagger Generator Based on Support Vector Machines. In Proceedings of the 4th International Conference on Language Resources and Evaluation, Lisbon, Portugal, 26–28 May 2004, pp. 43--46.

Habash N. and Rambow (2005).Part-of-Speech Tagging and Morphological Disambiguation in One Fell-Swoop.In:Proc. of the American Association of Computational Linguistic

Conference (ACL) Short Papers, Michigan,USA Schmid H. (1994). Proceedings of International Conference on New Methods in Language Processing, Manchester, UK.

Jurafsky D. and Martin J.H. (2009). Speech and Language Processing: An Introduction to Natural Language Processing, computational linguistics, and speech recognition, 2nd Ed. New Jersey: Prentice Hall.

Khoja S., Garside R. and Knowles G. (2001). A Tagset For The Morphosyntactic Tagging Of Arabic.In Proceedings of Corpus Linguistics. Lancaster, UK, pp 341-353.

Kudo T. and Yuji Matsumoto Y. (2000). Use of Support Vector Learning for Chunk Identification.In Proceedings of CoNLL-2000 and LLL-2000.

Kurcera H. and Francis W. N. (1967). Computational Analysis of Present-Day American English.Brown University Press, Providence, RI.

Laabdelaoui R., Boumalk A., Iazzi, E.M., Souifi H. and Ansar K.(2012). Manuel de conjugaison de l'Amazighe. Publications de l'IRCAM.

Lafferty J., McCallum A. and Pereira F. (2001): Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. In Proceedings of ICML-01, pp. 282-289.

Outahajala M., Zenkouar L. and Rosso P. (2014). Construction d'un grand corpus annoté pour la langue Amazighe. La revue Etudes et Documents Berbères n°33, pp.57-74.

Outahajala M. (2015). Apprentissage supervisé d'un étiqueteur morphosyntaxique automatique de la langue Amazighe. Thèse de Doctorat. Ecole Mohammedia d'Ingénieurs, Université Mohamed V-Rabat

Ratnaparkhi A., Reynar J. and Roukos S. (1994). A Maximum Entropy Model for Prepositional Phrase Attachment. In Proceedings of the Human Language Technology Workshop (ARP, 1994), pages 250-255.

Rachidi A. and Mammass D. (2005). Vers un système d'écriture informatique Amazighe : méthodes et développements, RECITAL 2005.

Diab M., Hacioglu K. and Jurafsky D. (2004). Automatic Tagging of Arabic Text: From Raw Text to Base Phrase Chunks. HLT-NAACL, 149-152.

Khoja S. (2001). APT: Arabic Part-of-speech Tagger. Workshop NAACL.

Brill E. (1993). Tagging an unfamiliar text with minimal human supervision. In proceedings of the Fall Dymposium on Probabilistic Approch to Natural Language.

Vers la construction des ressources linguistiques nécessaires pour la génération de la langue amazighe à partir de l'inter-langue UNL

Imane Taghbalout[1] **Fadoua Ataa Allah**[2] **Mohamed El Marraki**[1]
[1] LRIT, Faculté des sciences, Université Mohammed V, Rabat, Maroc {taghbalout.imane, elmarrakimohamed}@gmail.com
[2] CEISIC, Institut Royal de la Culture Amazighe, Rabat, Maroc ataaallah@ircam.ma

Résumé

La traduction automatique multilingue à base d'une inter-langue a été largement considérée comme l'approche de traduction la plus attrayante dans le cas des langues peu dotées informatiquement. Dans cet article, il s'agit de l'inter-langue UNL (Universal Networking Language, langage de réseau universel) qui permet à tout texte en langue source à être traduit vers l'ensemble des langues cibles participantes au projet UNL, et cela par la conversion du sens porté par le texte source en un graphe UNL, et la déconversion de ce graphe en la langue cible. Ce processus d'enconversion et de déconversion nécessitent la préparation d'un ensemble de ressources linguistiques, à savoir : une base lexicale et une base des règles grammaticales. Dans cet article, nous décrivons les ressources linguistiques nécessaires pour la déconversion du language UNL à la langue amazighe.

1. Introduction

L'informatisation de la langue amazighe est un enjeu stratégique garantissant sa survie et son positionnement dans la société de l'information. Dans ce sens, plusieurs efforts ont été déployés durant cette dernière décennie pour doter cette langue de ressources linguistiques et d'outils du Traitement Automatique des Langues (TAL). Cependant, à notre connaissance, il n'existe pas de travaux sur la Traduction Automatique (TA) de la langue amazighe. Pour cette raison, nous avons entamé les premières étapes de réalisation d'un Système de Traduction Automatique (STA). Certes l'approche statistique est l'approche la plus prometteuse dans le domaine de la TA mais elle requiert des corpus de très grande taille. Or, pour notre cas, la langue amazighe est une langue peu dotée informatiquement, il sera difficile de trouver un corpus de taille supérieure à quelques milliers de textes. Pour cette raison, nous avons opté pour l'approche linguistique : la traduction via l'inter-langue UNL. Dans ce cas, la traduction de n'importe quelle langue source vers n'importe quelle langue cible est le processus qui consiste à « convertir » la phrase source vers la représentation UNL puis à « déconvertir » la phrase cible à partir de cette représentation UNL. Le choix du langage UNL comme une langue pivot est basé premièrement sur le fait que l'UNL est conçu non seulement pour les langues les plus avancées informatiquement mais aussi pour les langues peu dotées et qui sont

en voie de disparition, deuxièmement parce que ce langage donne la possibilité de travailler dans un environnement multilingue. Du coup, la réalisation du convertisseur Amazighe-UNL et du déconvertisseur UNL-Amazighe permettra de traduire tout texte amazighe vers les différentes autres langues participantes dans le projet UNL.

Dans la deuxième partie de cet article, nous présentons brièvement le projet UNL. Dans la partie qui suit, nous aborderons le processus de construction du dictionnaire UNL-Amazighe, en l'occurrence l'identification et la formalisation des paradigmes flexionnels et les cadres de sous-catégorisation amazighes. La quatrième partie sera consacrée à l'implémentation des règles grammaticales de génération du texte amazighe à partir de la représentation sémantique UNL.

2. Le projet UNL (Universal Networking Language)

L'organisation des Nations Unis a lancé le projet UNL (Universal Networking Language) sous les auspices de l'institut des études avancées de l'université des Nations Unis de Tokyo en 1996 (Uchida *et al.*, 1999). Le but de ce projet est de permettre à toute personne du monde entier d'accéder à toutes les informations existantes sur Internet dans sa langue maternelle, favorisant ainsi le multilinguisme et réduisant les contraintes d'accéder à l'information à cause des barrières linguistiques. Pour cela, l'équipe de ce projet a développé un langage formel, appelé UNL, qui permet de coder le sens d'une information sous la forme d'un graphe qui se compose d'un ensemble de nœuds reliés par des arcs. Chaque nœud contient un "UW", Mot Universel (Universal Word en anglais) et chaque arc porte une relation sémantique entre deux nœuds. UWs sont souvent accompagnés d'un ensemble de propriétés grammaticales appelées des attributs (Uchida and Zhu, 2004). Les définitions de chacune de ces éléments de base de l'inter-langue UNL sont :

Mots Universaux (UWs) : constituent le vocabulaire du langage UNL, se sont des mots anglais accompagnés d'un ensemble de restrictions sémantiques et linguistiques.

Attributs Universaux : représentent les propriétés grammaticales qui peuvent enrichir la description des mots universaux. Par exemple, le mot universel "UW" qui correspond au mot anglais 'play' est 'play (icl>do)'. (icl>do) est ajouté pour dire qu'il s'agit d'un verbe. Si le verbe 'play' est conjugué au passé, l'attribut '@past' doit être ajouté à l'UW 'play (icl>do)'. Ainsi, nous obtenons la syntaxe suivante de l'UW : 'play (icl>do, @past)'.

Relations Universaux : sont des relations syntactico-sémantiques binaires qui connectent une paire de nœuds dans un graphe UNL. Le système UNL définit un ensemble de labels pour ces relations suivant leurs rôles. Par exemple, la relation "agt" (agent) définit la chose ou la personne qui initie une action.

La traduction automatique via UNL fait appel à un ensemble de ressources linguistiques et d'infrastructures techniques.

2.1. Ressources linguistiques

Les ressources linguistiques sont stockées dans le framework « UNLarium », qui consiste en des bases de données lexicales (dictionnaires), des bases de règles (grammaires) et des bases de documents (des corpus).

- Dictionnaire UNL qui liste les UWs avec leurs propriétés linguistiques et sémantiques dans un ordre alphabétique.
- Dictionnaire LN (Langue Naturelle) qui liste les entrées lexicales des langues naturelles avec leurs propriétés linguistiques.
- Dictionnaire LN-UNL, c'est un dictionnaire bilingue qui relie les entrées lexicales d'une langue naturelle à leurs correspondants en UNL. Ce dictionnaire peut être exploité de deux manières : Soit sous une forme générative, dans laquelle le dictionnaire comporte seulement les formes de base (lemmes). Dans ce cas, il est appelé dictionnaire de génération. Soit sous une forme énumérative dans laquelle le dictionnaire LN-UNL liste toutes les formes fléchies d'un lemme donné. Et dans ce cas, il est appelé dictionnaire d'analyse. Il est exploité principalement dans la phase d'analyse des langues naturelles.
- Une base de connaissance UNL qui regroupe toutes les relations possibles entre les mots universaux, une liste des règles grammaticales responsables de la conversion des textes en langues naturelles et la déconversion des graphes UNL.
- Grammaire UNL-NL sont l'ensemble de règles qui convertissent les phrases en des graphes UNL et vice-versa. Il existe deux types de règles : règles transformationnelles utilisées pour la génération des phrases en langue naturelle à partir du graphe UNL et vice-versa. Et règles de désambiguïsation utilisées pour améliorer la performance des règles transformationnelles en limitant leur application.

Le schéma suivant illustre l'architecture de fonctionnement des différentes ressources linguistiques que nous avons expliquées ci-dessus.

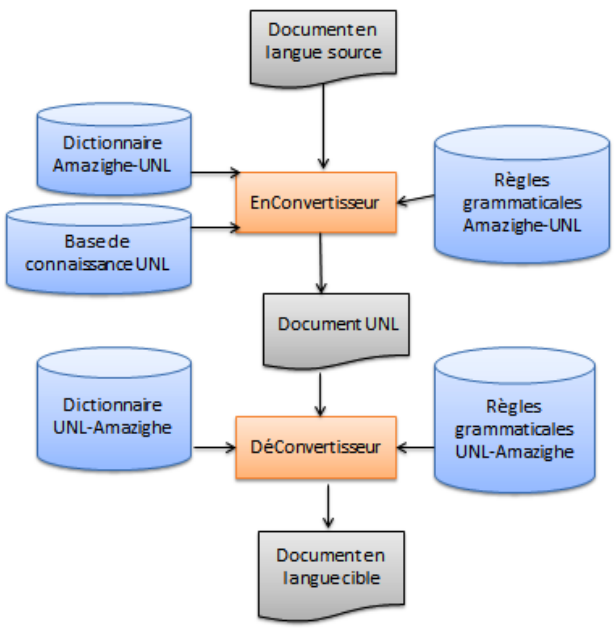


Figure 1 : Processus de l'Enconversion et de la Déconversion en UNL

2.2. Infrastructures techniques

En plus des ressources linguistiques, la réalisation d'un système de traduction d'une langue via UNL se base sur deux composantes logicielles : un système d'analyse (convertisseur) appelé IAN (Interactive Analysis System) et un système de génération (déconvertisseur) des langues naturelles, appelé EUGENE (dEep-to-sUrface natural language GENErator). Les processus d'analyse et de génération se déroulent en trois étapes sur le fichier d'entrée :

- Segmentation : la division du document d'entrée en une série de graphes isolés par le système de génération, et en des phrases par le système d'analyse.
- Tokenisation : l'identification des UWs, des relations universelles et des attributs universels, par le système de génération de chaque graphe du document d'entrée ; et l'identification des unités lexicales par le système d'analyse.
- Transformation : l'application des règles grammaticales de transformation sur chaque graphe, par le système de génération, afin de le convertir en une phrase en langage naturel, et sur chaque phrase, par le système d'analyse, afin de générer le graphe qui lui correspond.

IAN opère d'une manière semi-automatique, car la désambiguïsation du sens d'un mot reste une tâche humaine. En effet, lorsque le système d'analyse, lors de la phase de tokenisation, identifie un mot supportant plusieurs sens, dans ce cas l'utilisateur intervient pour fixer le sens adéquat. Par exemple le mot « cuisinière » veut dire à la fois l'appareil qui fait chauffer la nourriture, et la personne qui sait faire la cuisine. Cependant, EUGENE opère automatiquement sans aucune intervention humaine Construction du dictionnaire UNL-Amazighe.

En vue de bien créer un dictionnaire soit d'analyse ou de génération, nous devrons se disposer d'un ensemble de caractéristiques linguistiques telles que la catégorie grammaticale, la transitivité, le paradigme flexionnel et le cadre de sous-catégorisation auxquels appartient l'entrée lexicale.

Chaque entrée du dictionnaire UNL-Amazighe a le format suivant : (Teixeira Martins et Avetisyan, 2009) :

[NLW] {ID} 'UW' (ATTR …) < FLG, FRE, PRI >;

où:

- NLW : l'entrée (Mot amazighe).
- ID : Identifiant de l'entrée.
- UW : (Universal Word) Mot universel.
- ATTR : la liste des traits morphologiques (par exemple : paradigmes flexionnelles), syntaxiques (cadres de sous-catégorisation) et sémantiques.
- FLG : code de la langue accordé par ISO 639-3 (Ber pour la langue Berbère).
- FRE : la fréquence d'occurrence du NLW dans un texte.
- PRI : la priorité du NLW lors de la génération de la langue.

La construction du dictionnaire ne se complète qu'avec la présence des traits linguistiques : paradigmes flexionnels et cadres de sous-catégorisation parmi la liste des attributs.

Dans la suite de cette partie, nous présentons les paradigmes flexionnelles et les cadres de sous-catégorisations amazighes que nous avons identifié et formalisé.

2.3. Formalisation des paradigms flexionnelles amazighes

La langue amazighe est une langue morphologiquement riche, son système flexionnel est complexe. En effet, la formation des mots fléchis amazighes fait appel soit à l'un de ces procédés : préfixation, suffixation ou infixation, soit à la combinaison de ces trois procédés.

Les mots amazighes sont classés en dix catégories lexicales : nom, verbe, adverbe, préposition, pronom, conjonction, interjection, numéral et particule (Boukhris *et al.*, 2008; Ataa Allah *et al.*, 2014a). Puisque, en amazighe, les conjonctions, les particules, les adverbes, les prépositions, les pronoms, et les interjections sont invariants, ces catégories ne subit aucun processus de flexion. Dans cette partie, nous présenterons les paradigmes flexionnelles des noms et des verbes amazighes que nous avons formalisés (Taghbalout *et al.*, 2016).

2.3.1. Paradigmes flexionnelles nominales

La construction des paradigmes flexionnels de la catégorie nominale a présenté pour nous un vrai challenge, étant donné qu'il existe plusieurs formes de pluriels et qu'il y a un manque de travaux sur la classification des noms amazighes par rapport à la forme du pluriel, l'état d'annexion et la forme du féminin.

En se basant sur des heuristiques et sur les travaux de (Boukhris *et al.*, 2008; Nejme *et al.*, 2012; Raiss et Cavalli-Sforza, 2012), nous avons pu construire, dans un premier niveau (Fig. 2), des classes lexicales des noms ayant la même forme du pluriel. Dans un deuxième niveau, nous avons créé des sous-classes pour chaque classe construite dans le niveau 1 mais cette fois suivant la forme de l'état d'annexion et du genre. Il est à noter qu'un nom amazighe peut appartenir à plusieurs classes à la fois selon sa variété régionale.

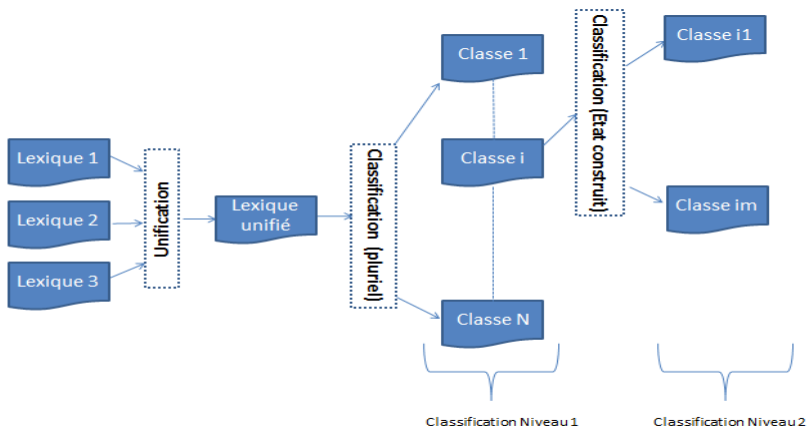


Figure 2 : Méthode de création des classes de noms amazighes

En procédant de cette manière, nous avons pu identifier 90 classes nominales, et du coup, nous avons formalisé 90 paradigmes flexionnels (Taghbalout *et al.*, 2015). La table ci-dessous présente, à titre d'exemple, les règles flexionnelles du paradigme M49 dont appartient le nom «ⴰⵙⵏⴰⵙ» [Instituteur].

Règles flexionnelles	Explication	Forme fléchie
MCL&SNG&NOM: =0>»»;	Pas de changement dans le cas où le nom est au masculin, singulier, et à l'état libre	ⴰⵙⵏⴰⵙ
MCL&PLR&NOM:=»ⵉ»<1,0>»ⵍ»;	Changement de la première lettre par "ⵉ" et suffixation de "ⵍ" lorsque le nom est au masculin, pluriel, et à l'état libre	ⵉⵙⵏⴰⵙⵍ
MCL&SNG&CTS: = «ⵓ»<1;	Changement de la première lettre par "ⵓ" lorsque le nom est au masculin, singulier, et à l'état d'annexion	ⵓⵙⵏⴰⵙ
MCL&PLR&CTS: =»ⵉ»<0 ,1>»ⵍ»;	Changement de la première lettre par "ⵉ" et suffixation de "ⵍ" lorsque le nom est au masculin, pluriel, et à l'état d'annexion	ⵉⵙⵏⴰⵙⵍ
FEM&SNG&NOM:=»ⵜ»<0,0>»ⵜ»;	Préfixation de la lettre "ⵜ" et suffixation de la lettre "ⵜ" lorsque le nom est à l'état libre, au féminin, et au singulier	ⵜⴰⵙⵏⴰⵙⵜ
FEM&SNG&CTS: =»ⵜ»<1, 0>»ⵜ»;	Changement de la première lettre par "ⵜ" et suffixation de "ⵜ" lorsque le nom est à l'état d'annexion, au féminin, et au singulier	ⵜⵙⵏⴰⵙⵜ
FEM&PLR&NOM:=»ⵜⵉ»<1,0>»ⵉⵍ»;	Changement de la première lettre par "ⵜⵉ" et suffixation de "ⵉⵍ" lorsque le nom est à l'état libre, au féminin, et au pluriel	ⵜⵉⵙⵏⴰⵙⵉⵍ
FEM&PLR&CTS:=»ⵜ»<1,0>»ⵉⵍ»;	Changement de la première lettre par "ⵜ" et suffixation de "ⵉⵍ" lorsque le nom est à l'état d'annexion, au féminin, et au pluriel	ⵜⵙⵏⴰⵙⵉⵍ

Tableau 1 : Règles flexionnelles du paradigme M49

Légende :

- *"a"<0;= Préfixation du caractère "a";*
- *0>»a»;= Suffixation du caractère "a";*
- *MCL : Masculin; FEM : Féminin;*
- *CTS : Etat d'annexion ;*
- *NOM : Etat libre.*

2.3.2. Paradigmes flexionnels verbales

La génération flexionnelle du verbe amazighe donne lieu à 46 formes fléchis. Le mode participiale (PTP) renvoie 4 formes, le mode impératif (IMP) renvoie 6 formes et le mode indicatif (IND) renvoie 36 formes (9 formes distinctes pour chacun des quatre aspects suivants : aoriste, accompli, accompli négatif et l'inaccompli).

Pour formaliser les paradigmes flexionnelles de la catégorie verbale, nous nous sommes basés sur la classification des verbes proposée par (Laabdelaoui *et al.*, 2012) et par (Ataa Allah et Boulaknadel, 2014b), selon cette classification, les verbes sont classés en 31 classes selon les oppositions aoriste/accompli et aoriste/inaccompli. Notre méthodologie de classification vise à extraire de nouvelles sous-classes de l'une de ces 31 classes de telle façon, chaque sous-classe rassemble tous les verbes ayant les même règles morphotactiques et morphonologiques de génération de toutes leurs formes conjuguées. Le schéma ci-dessous illustre notre méthodologie de création des sous-classes verbales.

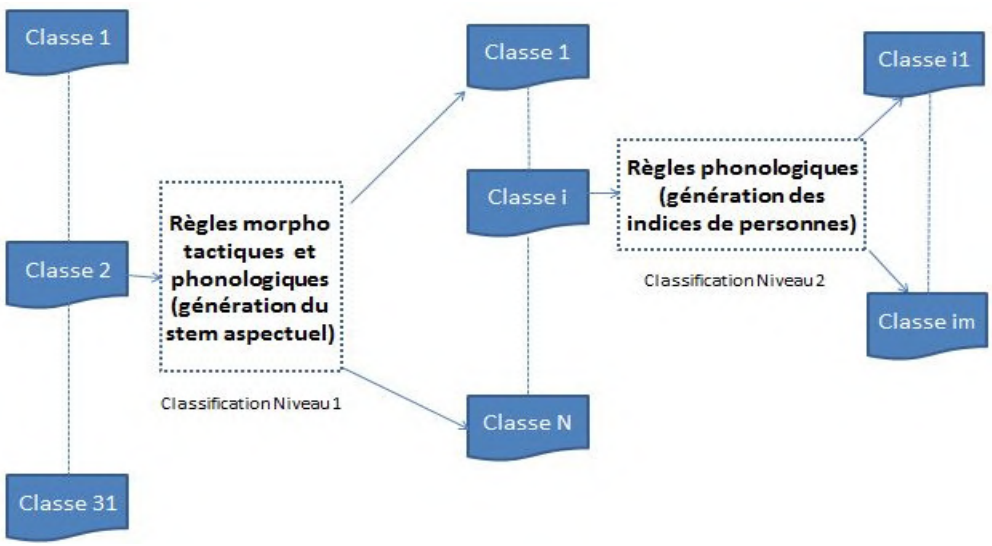


Figure 3 : Processus de création des classes verbales amazighes

A titre d'exemple, nous avons fait sortir à partir de la classe N° 2, trois sous-classes 2-1, 2-2, et 2-3. Les verbes appartenant à ces classes ne partagent pas les mêmes règles morphotactactiques de génération des formes aspectuelles Accompli négatif et Inaccompli.

N° Classe	Aspect	Procédé morphotactique de génération des aspects
2-1	Accompli négatif	Infixation pré-finale de ⵅ
	Inaccompli	Préfixation de ++
2-2	Accompli négatif	Pas de changement
	Inaccompli	Préfixation de ++, et dégémination
2-3	Accompli négatif	Pas de changement
		Préfixation de ++

Tableau 2 : Les sous-classes de la classe 2

Le nombre de classes verbales que nous avons pu formalisé jusqu'à maintenant est 58 paradigmes flexionnels verbales (Taghbalout *et al.*, 2015b). Il est à noter qu'un verbe amazighe peut appartenir à plusieurs classes à la fois selon le sens qu'il porte et aussi selon sa variété régionale.

La table « tab.3 » présente un extrait des règles flexionnelles responsables de la génération de la conjugaison du verbe ' ' (arriver) à l'accompli négatif (PFV&NEG).

Verbe	Règles flexionnelles UNL	verbe conjugué
'' [awd] *arriver*	1PS&PFV&NEG&IND := "ⵖ"<1, "ⵅ"< [-1], 0> "ⵢ";	
	2PS&PFV&NEG&IND:= "ⵖ"<1,"ⵅ"< [-1], "+"<0, 0>"ⴰ";	
	3PS&MCL&PFV&NEG&IND:= "ⵖ"<1,"ⵅ"<[-1] , "ⵣ"<0;	
	3PS&FEM&PFV&NEG&IND:="ⵖ"<1,"ⵅ"<[-1], "+"<0;	
	1PP&PFV&NEG&IND:= "ⵖ"<1,"ⵅ"<[-1],"ⵍ"<0;	
	2PP&MCL&PFV&NEG&IND:= "ⵖ"<1,"ⵅ"<[-1], "+"<0,0>"ⵛ";	
	2PP&FEM&PFV&NEG&IND:= "ⵖ"<1,"ⵅ"<[-1], "+"<0, 0>"ⵛ+";	
	3PP&MCL&PFV&NEG&IND:= "ⵖ"<1, "ⵅ"<[-1], 0>"ⵍ";	
	3PP&FEM&PFV&NEG&IND:= "ⵖ"<1,"ⵅ"<[-1],0> "ⵍ+";	

Tableau 3 : Extrait des règles flexionnelles du paradigme M154

Légende :

- *1PP : 1ère personne du pluriel ; PFV : Accompli ; IND : le mode indicatif ; FEM : féminin ; MCL : masculin ;*
- *“a”<0;= Préfixation du caractère “a”;*
- *“a”<1;= Substitution de l’initial par le caractère “a” ;*
- *0>»a»;= Suffixation du caractère “a”.*

Dans le formalisme UNL, l’ordre d’apparition des règles est important. Comme le décrit la table ci-dessus, chaque règle flexionnelle exprime en premier lieu les règles morphotactiques pour avoir le radical aspectuel, ensuite les règles morpho-tactiques pour générer les indices de personnes suivant le mode et finalement les règles morphophonologiques. Toutes ces règles sont combinées, l’une après l’autre d’une manière linéaire pour générer la flexion désirée.

2.4. Formalisation des cadres de sous-catégorisation amazighes

La sous-catégorisation définit le nombre et le type d’arguments syntaxiques (spécificateur, complément, modificateur, adjoint...) qui coexistent avec la forme de base (le constituant) pour constituer un syntagme. Jusqu’à présent, nous avons identifié et formalisé 22 cadres de sous-catégorisation à savoir des cadres de sous-catégorisation verbales, prépositionnelles, adverbiales, … (Taghbalout *et al.*, 2015c).

Cadre de sous-catégorisation (Formalisation de l’UNL)	**Explication**	**Exemple**
VS (NP, ANM);	Les verbes admettant des sujets animés	ⵙⵀⵍⴰ ⵜⵕⵜ ! ‘écoute-moi’
VC (VH([ⴰⴰ]));	Les verbes modaux amazighes d’obligation admettant un syntagme verbal comme complément précédé par ⴰⴰ	ⵜⵍⴰⵣⵎ ⴰⴰ ‘il faut que’ ⵜⵍⴰⵣⵎ ⴰⴰ ⵕⴰⵣⵍ ⵜⴰⵓⵜⵎ ⵜ ⵙⴰⵙⴰⵙ ‘Il doit envoyer de l’argent à son père’
AC (PH ([l]));	Les adverbes admettant un complément précédé par la préposition ‘ⵍ’	ⵙⵇⵇⴰ ⵍ ‘à l’extérieur de’ ⵜⵀⵀⵖ ⵙ ⵙⵇⵇⴰ ⵍ ⵍⵎⴰⵜⵜⵍ ‘Il est allé à l’extérieur de la ville’
PC (NP, NOM);	Les prépositions admettant un syntagme nominale à son état libre comme complément	ⵙⵍⴰ ‘sans’ ⵜⵛⵛⴰ ⵙⵍⴰ ⴰⵎⵕⴰ ‘il a mangé gratuitement’
PS (PH([ⵙ]));	Les syntagmes prépositionnels admettant la préposition ‘ⵙ’ comme spécificateur	ⵍⴰⴰⵙ, ⵏⵉ, ⴰⴰⵔ, ⵙ ⵍⴰⵎⵍ ‘avec de l’eau’

Tableau 4 : Exemples de cadre de sous-catégorisation

3. Construction des règles de transformation UNL-Amazighe

Le processus de génération des phrases amazighes à partir des graphes sémantiques UNL fait appel à un ensemble de règles grammaticales, dites, règles de transformations. Les phrases amazighes générées et les graphes UNL sont supposés porter la même quantité d'informations en des structures différentes. La première structure arrange les informations en une liste de mots, alors que la deuxième les organise en un hyper-graphe. Ainsi, nous pouvons dire que la traduction depuis une langue naturelle vers UNL et depuis UNL vers une langue naturelle est une question de transformation des listes en des réseaux et vice-versa. L'application web EUGENE, conçue pour la génération, suppose que cette transformation doit être effectuée progressivement à travers la structure de données transitoire « arbre » qui pourrait venir entre la structure réseau et la structure liste.

Le dictionnaire UNL-Amazighe et la liste des règles de transformation UNL-amazighe sont deux fichiers séparables que nous chargeons sur l'outil de déconversion EUGENE pour pouvoir générer des phrases amazighes à partir de tout document UNL. Nous illustrons dans ce qui suit le processus de déconversion du document UNL suivant :

[S: S#1]
{org}
He gave her a book
{/org}
{unl}
agt(give :03.@past, 00 :01.@3.@male)
adr(give :03.@past, 00:05.@3.@female)
obj (give :03.@past, book :07.@3.@indef)
{/unl}
[/S]

L'exemple ci-dessus présente le cas d'une simple phrase UNL, elle comporte trois relations universelles obj, agt, et adr :

- obj : indique la chose qui est affectée directement par un évènement ou par un état.
- agt : indique la chose qui initie une action.
- adr : indique la personne recevant quelque chose (complément d'objet indirect).

Les attributs qui sont attachés aux UWs sont :

- @past : le temps passé de l'UW give(icl>do) 'donner'.
- @3 : la troisième personne.
- @male : le genre masculin.
- @female : le genre féminin.
- @indef : l'article accompagné au nom est indéfini.

Le graphe correspondant à la phrase UNL ci-dessus est :

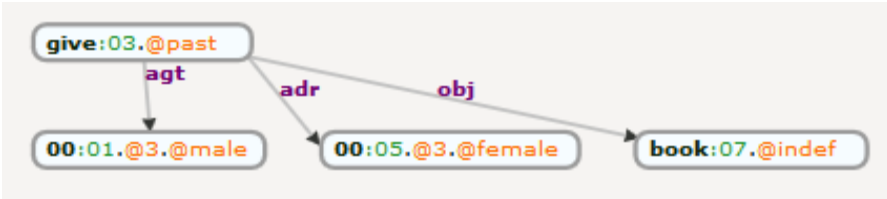


Figure 4 : Graphe UNL correspondant

Le processus de génération commence par parcourir le dictionnaire UNL-Amazighe pour chercher et extraire les mots amazighes équivalents aux mots universels du graphe UNL. Dans le cas du graphe 'Fig. 4', voici les entrées du dictionnaire extraites :

[ⴰⵏⵉⵙ]{}»book»(LEX=N,POS=NOU,LST=WRD,GEN=MCL,NUM=SNG,CAS=NOM, PAR=M6,FRA=Y0) <Ber,0,0>;

[ⵅⵔ]{}»give»(LEX=V, POS=VER, TRA=NTST, PAR=M162, FRA=Y0) <BER,0,0>;

Après ce stade du mapping lexical, vient la phase d'application des règles de transformations UNL-amazighe adéquates sur ces entrées lexicales pour prendre en considération lors de la génération l'ordre syntaxique et la flexion morphologique.

Ainsi, la phrase amazighe générée à partir du graphe UNL « Fig.4 » est «ⵙⵅⵔⴰ ⴰⵙ ⴰⵏⵉⵙ» 'il lui a donné un livre'.

```
[S:S#12]
    {org}
        He gave her a book
    {/org}
    {ber}
        ⵙⵅⵔⴰ ⵜⵜ ⴰⵏⵉⵙ
    {/ber}
    {unl}
        agt(give:03.@past,00:01.@3.@male)
        adr(give:03.@past,00:05.@3.@female)
        obj(give:03.@past,book:07.@indef)
    {/unl}
[/S]
```

Figure 5 : La phrase amazighe générée à partir du graphe UNL 'Fig. 4'

Parmi les règles de transformations, elles existent certaines qui sont indépendantes de la langue et d'autres qui lui sont propres. Jusqu'à présent, nous avons pu implémenter 70 règles de transformation spécifiques à la langue amazighe. La table ci-dessous présente quelques exemples de règles grammaticales que nous avons implémentées.

Règles de transformation	**Explication**
agt(%x,V;%y,N): = VS(%x,+PER = %y, +GEN=%y, +NUM= % y;%y, +CAS= CTS);	La relation sémantique Agent « agt » entre le nœud %x de catégorie grammaticale verbale et un autre %y de catégorie nominale devient une relation syntaxique (VS) dont le nom % y est le spécificateur du verbe et prend la marque de l'état d'annexion (CTS), le genre et le nombre à partir du verbe.
obj(%x,V; %y, N) : = VC (%x; %y, -CAS, +CAS = NOM);	La relation sémantique Objet « obj » entre le nœud %x de catégorie grammaticale verbale et un autre %y de catégorie nominale devient une relation syntaxique VC dont le nom % y est le complément du verbe, qui reste à l'état libre (NOM)
rsn(%x,V;%y,N): = VA(%x;PC([⊙], LEX= P, POS= PRE; %y, -CAS, +CAS= CTS));	La relation sémantique de cause « rsn » est transformée en la relation syntaxique VA (Adjoint du verbe) dont le nom %y est la cause de l'action du verbe %x introduit par la préposition ⊙
pos(%x,N; %y, D) : = NS(%x; %y, DIS = AFT);	La relation sémantique de possession « pos » est transformée en une relation syntaxique NS (spécificateur de nom) dont le déterminant possessif se place immédiatement après le nom

Tableau 5 : Exemples de règles de transformation de génération (UNL –Amazighe)

Conclusion

Les ressources linguistiques nécessaires à la traduction automatique comprennent toujours un dictionnaire et des règles grammaticales. L'élaboration de celles-ci est un processus incrémental et long. Nous avons divisé notre projet de traduction automatique en deux modules : un module d'analyse et un module de génération. Le présent article décrit le module de génération. Dans un premier temps, nous avons abordé le processus de réalisation du dictionnaire de génération UNL-Amazighe, en l'occurrence la formalisation des paradigmes flexionnels et les cadres de sous-catégorisation. Et dans un second temps, nous avons discuté les règles grammaticales de génération UNL-amazighe. Actuellement, nous disposons d'un dictionnaire de 2600 lemmes et d'une base de règles contenant 70 règles de transformation. Nous sommes en train de préparer un corpus de test pour évaluer l'exactitude de la grammaire implémentée en calculant la F-mesure ; en parallèle nous continuons à alimenter notre dictionnaire et à élaborer de nouvelles règles de transformation UNL-amazighe.

Références

Ataa Allah, F., Boulaknadel, S., Souifi, H. (2014). 'Jeu d'Etiquettes Morphosyntaxiques de la Langue Amazighe', Asinag, n°9, pp. 171-184, ISSN : 2028-5663.

Ataa Allah, F. and Boulaknadel S. (2014). 'Amazigh Verb Conjugator'. 9th International Conference on Language Resources and Evaluation (LREC 2014), Reykjavik, Iceland, May 26-31 2014, pp. 1051-1055.

Boukhris, F., Boumalk, A., El Moujahid, E., Souifi, H. (2008). 'La nouvelle grammaire de l'amazighe', IRCAM, Rabat, Maroc.

Laabdelaoui R., Boumalk A., Iazzi, E.M., Souifi H., Ansar K. (2012). 'Manuel de conjugaison de l'amazighe' . IRCAM, Rabat, Morocco.

Nejme, F., Boulaknadel, S., Aboutajdine, D. (2012). 'Toward an Amazigh language processing', the 3rd Workshop on South and Southeast Asian Natural Language Processing, Mumbai, India.

Raiss, H. and Cavalli-Sforza, V. (2012). 'Amazigh Nouns Morphological Analyzer', 5ème Conférence internationnale sur les TIC pour l'amazighe, Rabat, Maroc.

Taghbalout, I., Ataa Allah, F., El Marraki, M. (2015). 'Amazigh Noun Inflection in the Universal Networking Language'. International Journal of Education and Information Technologies.

Taghbalout, I., Ataa Allah, F., El Marraki, M. (2015). 'Amazigh verb in the Universal Networking Language'. 12th ACS/IEEE International Conference on Computer Systems and Applications AICCSA, Marrakech, Morocco .

Taghbalout, I., Ataa Allah, F., El Marraki, M. (2015). 'Amazigh Representation in the UNL Framework: Resource Implementation'. the International Conference on Advanced Wireless Information & Communication Technologies. Procedia Computer Science.

Taghbalout, I., Ataa Allah, F., El Marraki, M. (2016). 'Towards UNL based machine translation for Amazigh language'. International Journal of Computational Science and Engineering;

Teixeira M. R. and Avetisyan V. (2009). 'Generative and Enumerative Lexicons in the UNL Framework', in proceedings of 7th International Conference on Computer Science and_Information Technologies, (CSIT 2009), Yerevan, Armenia.

Uchida, H., Zhu, M., Senta, T.D. (1999), 'UNL: A gift for a millinnium'. Institute of Advanced Studies, United Nations University, Tokyo.

Uchida, H. and Zhu, M. (2004) 'The Universal Networking Language (UNL) Specification Version 3.0'. Edition 3, Technical Report, UNU.

Annexe

Symbole	Explication
VS	Spécificateur du verbe
VC	Complément du verbe
AC	Complément de l'adverbe
NP	Syntagme nominal
PC	Comlément d'une préposition
ANM	Etre animé
NOM	Etat libre
NS	Spécificateur de nom
VH	La tête du syntagme verbale
PH	La tête du syntagme prépositionnel

Corpus multilingues pour les langues peu dotées

Nina Miftah[1] Fadoua Ataa Allah[2] Imane Taghbalout[1]
[1] LRIT, Faculté des Sciences, Université Mohammed V, Rabat, Maroc
[2] CEISIC, Institut Royal de la Culture Amazighe, Maroc
{ninam202, taghbalout.imane}@gmail.com
ataaallah@ircam.ma

Résumé

Les corpus électroniques constituent de nos jours un élément essentiel et une base référentielle élémentaire, présentant les faits d'une langue, pour bien mener des recherches linguistiques, philologiques et informatiques. Dernièrement, l'exploitation de corpus multilingues dans les études contrastives a connu un succès croissant et leur valeur ajoutée pour un éventail d'études sémantiques et syntaxiques, ainsi que pour des études terminologiques, a été soulignée dans de nombreuses publications.

Ainsi, dans le présent article nous abordons les corpus multilingues, et nous établissons une étude comparative entre ces corpus. Dans un second temps, nous présentons l'alignement des corpus multilingues.

1. Introduction

A l'époque de l'informatique, le nombre des documents électroniques augmentent de façon exponentielle et l'utilisation de ces documents joue un rôle très important dans de nombreuses disciplines. Le traitement automatique du langage (TAL), l'intelligence artificielle, la linguistique de corpus, la terminologie, les sciences humaines figurent parmi les disciplines qui se sont intéressées aux textes sur support électronique.

Dans nos jours, ces documents sont devenus importants pour les systèmes de traduction automatique (TA) pour certaines langues telles que l'anglais, français, italien, etc. Notamment, ils sont appliqués dans les différentes approches de la TA : des approches expertes fondées sur des règles linguistiques, des approches statistiques, ainsi que des approches hybrides (Miftah, 2016). Toutefois, la recherche sur la traduction automatique pour des langues dites « peu dotées » doit faire face au défi du manque de données.

Nous prenons comme exemple de langues peu dotées « la langue amazighe ». La question que se pose alors est : comment doter la langue amazighe pour acquérir la place des langues avancées ? Il s'avère primordial de la doter de ressources linguistiques et de corpus riches et exploitables essentiel à son traitement automatique. Dans ce contexte, nous nous s'intéressons dans ce travail à l'étude des corpus multilingues au profit de la traduction automatique.

Les corpus sont des collections de données sélectionnées et organisées selon des critères explicites pour servir d'échantillon pour un traitement particulier ou de référence pour fournir une information en profondeur. Généralement, les corpus sont caractérisés par la nature de la langue traitée et par leur contenu. Ainsi, un corpus peut représenter une langue ou plusieurs, comme il peut contenir du texte ou du multimédia. (Ataa Allah, 2015). Principalement, nous distinguons deux types de corpus multilingues : les corpus parallèles et les corpus comparables sans oublier les corpus alignés. Les corpus parallèles sont composés de paires de documents qui ont des traductions mutuelles alors que les corpus comparables sont composés de documents ayant des traits communs tels que le genre, la période, le domaine, etc, sans être des traductions, et enfin les corpus alignés qui peuvent constituer des ressources de base pour les recherches linguistiques comparatives et l'étude théorique de la traduction.

Afin de mieux élaborer les corpus, il est nécessaire d'en connaître plus à fond les caractéristiques. Nous nous concentrons sur les corpus conçu à la traduction automatique. Dans la suite de cet article, nous présentons, dans la section 2, les corpus multilingues. Puis, nous détaillons, dans la section 3, les corpus parallèles et les corpus comparables. Nous passons, ensuite, aux aligneurs, en présentant les méthodes d'alignement classiques qui sont à la base de la plupart des outils d'exploitation des corpus disponibles dans la littérature, dans la section 4. Alors que nous consacrons la section 5 à la conclusion.

2. Les corpus multilingues

Les corpus peuvent être qualifiés de monolingues ou multilingues, selon les langues traitées. Les monolingues contiennent des textes en une seule langue, tandis que les corpus multilingues contiennent des textes en plusieurs langues sélectionnées selon les mêmes critères. Les corpus multilingues sont des corpus composés de documents dans des langues différentes.

Les informations qui peuvent être mises à jour par l'investigation et l'analyse de ces corpus font une ressource importante pour la traduction automatique. En extraction lexicale, ces corpus permettent de suivre automatiquement l'évolution d'une langue. La section suivante est consacrée à la description des différents corpus multilingues et à la présentation des caractéristiques de chacun entre ces derniers. Deux types de corpus multilingues ont été définis dans la littérature, nous distinguons les corpus parallèles et les corpus comparables.

3. Les corpus parallèles et comparables

3.1. Corpus parallèle

Les corpus parallèles sont constitués de paires de textes source (original) et cible (traduction). Les textes parallèles existent depuis fort longtemps, avant l'apparition de l'informatique. Et ceux à travers les recherches anthropologiques qui ont découvert dans l'Afrique du Nord, des objets (stèles) portant des inscriptions en tifinagh, les plus anciennes remontent à plus de 5.000 ans. La plus connue est celle de « Dougga »[1] elle comportait des textes parallèles punique et

1 https://fr.wikipedia.org/wiki/Dougga

amazighe. Egalement, en juillet 1799, les soldats de l'armée de Napoléon découvraient près de la ville de Rosette, sur le delta du Nil, en Egypte, « la pierre de Rosette »[2].

L'histoire des corpus parallèles a commencée par l'étude de ce fragment, sur lequel figuraient trois systèmes d'écriture (hiéroglyphe, démotique et grec) qui a permis à Champollion de déchiffrer l'écriture hiéroglyphique (Le Serrec, 2008). A part ces deux exemples, nous retrouvons de nombreux types de textes parallèles que ce soit des traités, des contrats, des textes sacrés, des ouvrages littéraires ou des documents scientifiques. L'informatique ayant facilité l'exploitation des textes parallèles, et du coup, plusieurs travaux les ont utilisés notamment en traduction automatique, vu qu'ils constituent une mine de solutions de traduction réutilisables (Le Serrec, 2008).

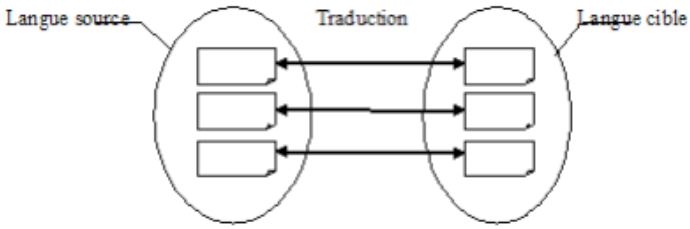


Figure 1. Schéma d'un corpus parallèle

En 1996. Sinclair a donné une définition globale des corpus parallèles :

« A parallel corpus is a collection of texts, each of which is translated into one or more other languages than the original. The simplest case is where two languages only are involved: one of the corpora is an exact translation of the other » . (Rappazzo, 2014).

L'équivalent de cette définition anglaise est « Un corpus parallèle est un recueil de textes, dont chacun est traduit en une ou plusieurs autres langues que l'original. Le cas le plus simple est d'avoir juste deux langues qui sont impliquées : l'un des corpus est la traduction exacte de l'autre ».

En fait, un corpus parallèle est constitué d'un ensemble de couples de documents tel que, pour un couple, un des documents est la traduction de l'autre. Les exemples les plus connu de corpus parallèles multilingues sont le corpus Europarl et le corpus Hansard (Harastani. 2014).

Afin d'être utile, un corpus parallèle doit receler des exemples de traductions similaires aux phrases constituant le nouveau texte à traduire dans le domaine de la traduction automatique (TA) ou de la traduction assistée par ordinateur (TAO). La valeur d'un corpus parallèle croît donc avec sa taille : plus un tel corpus est massif, plus on a de chance d'y trouver l'information cherchée. Par conséquent, plus un corpus parallèle contient de langues, plus il permet de traiter un grand nombre de couples, ce qui le rend d'autant plus intéressant (Authoul, 2012).

2 https://fr.wikipedia.org/wiki/Pierre_de_Rosette

L'information textuelle seule n'est pas suffisante pour la plupart des applications du TAL dans un corpus parallèle. En effet, la mise en correspondance entre les différentes parties équivalentes, à un niveau de granularité suffisant - par exemple les paragraphes ou les phrases - est indispensable pour mettre en œuvre les applications utiles. Cette information de correspondance est appelée alignement (Authoul, 2012).

C'est seulement à partir des années 1980 que les textes parallèles ont commencé à être exploités de façon systématique dans le cadre du traitement automatique des langues. Parmi les corpus parallèles de référence nous pouvons citer (Hazem et Emmanuel, 2013) :

-Le corpus Hansard : créé dans les années 1980. Ce corpus est composé de textes anglais et français extraits des transcriptions des débats du parlement canadien de 1970 à 1988.

-Le corpus Europral : ce corpus rassemble des textes du parlement européen dans 11 langues, avec plus de 20 millions de mots par langue.

-Le corpus de Bible : construit dans les années 1999, ce corpus a été organisé en 13 langues.

-Le corpus JRC-ACQUIS : ce corpus est constitué de textes des acquis communautaires de l'Union Européenne (EU) en majorité légiste dans 20 des langues officielles de l'Union Européenne.

Les corpus parallèles constituent un élément moteur pour la construction de lexiques bilingues robustes, la traduction automatique et la recherche d'information interlingue. Ils sont, généralement, construits par des traducteurs humains, qui à leur tour font appel à un lexique bilingue existant pour guider la traduction des textes. Néanmoins, ces corpus sont par nature des ressources rares notamment pour des domaines spécialisés, et difficile à constituer par rapport aux corpus comparables.

3.2. Corpus comparable

Au vu des différents problèmes dont souffrent les corpus parallèles, notamment du fait de leur rareté, et encore plus dans des domaines de spécialités et pour des couples de langues peu dotées, il devenait nécessaire d'y trouver une alternative. C'est en 1995 que fut l'introduction des méthodes permettant d'aligner des corpus non parallèles (corpus parallèles bruités puis corpus comparable). (Bouamor, 2014)

Les corpus comparables représentent ainsi une sous-catégorie de corpus multilingues. Généralement, dans la littérature, la notion de « corpus comparable » est assez vague. En fait, la communauté de chercheurs considère qu'un corpus comparable contient des documents qui ne sont pas des traductions l'un de l'autres, mais « étroitement liés par les mêmes contenus » aux « niveaux de parallélisme différents, tels que des mots, des chaînes de mots, des phrases, etc. » (Thi Ngoc Diep, 2012).

Nous trouvons plusieurs manières de définir un corpus comparable. Que ces définitions soient basées sur des critères communicationnels, lexicaux ou linguistiques, elles restent peu précises. Les corpus comparables sont composés de documents ayant des traits communs tels que le genre, la période, le domaine, etc, sans être des traductions. (Rebout, 2012).

Bowker et Pearson (2002) les définissent :

« [...] comparable corpora consist of sets of text in different languages that are not translations of each other. We use the word 'comparable' to indicate that the text in different languages have been selected because they have some characteristics or features in common; the one and only features such as that distinguishes one set from another in a comparable corpus is the languages in which the text are written ».

L'équivalent de cette définition anglaise est «[...] Corpus comparables sont constitués d'ensembles de textes dans différentes langues qui ne sont pas des traductions de l'autre. Nous utilisons le mot «comparable» pour indiquer que les textes ont des caractéristiques communes ».

En effet, il s'agit de deux ensembles de textes sélectionnés selon les mêmes critères, qui diffèrent, dans la plupart des cas, uniquement par la langue, mais qui ne sont pas des traductions mutuelles. Le terme comparable indique que ces textes partagent certaines caractéristiques ou certains traits, par exemple dans la majorité du temps, nous sélectionnons des textes qui ont en commun : le sujet, la période ou le degré de technicité, etc (Harastani, 2014).

Li et Gaussier (2010) ont par la suite proposé une mesure de comparabilité quantitative définie comme la proportion du vocabulaire pour lequel il existe une traduction dans le corpus. Ils montrent que la qualité des dictionnaires extraits de corpus comparables est directement reliée à ce degré de comparabilité (Rebout, 2012).

Deux caractéristiques des corpus comparables ont été plus particulièrement discutées dans la littérature de l'alignement bilingue : la taille et le degré de comparabilité d'un corpus comparable.

Taille de corpus : La taille d'un corpus est un compromis entre le but de l'étude et le temps que l'on a à disposition. Un corpus ne doit pas forcément être énorme, ce qui compte c'est qu'il doit bien représenter la langue étudiée. Evidemment, un corpus de petite taille ne va pas représenter suffisamment toutes les caractéristiques linguistiques de la langue (Rappazzo, 2014).

Comparabilité de corpus : Afin de quantifier dans quelle mesure les textes d'un corpus sont comparables, nous nous basons sur la notion de degré de comparabilité d'un corpus. Un corpus a un degré de comparabilité élevé s'il comprend des textes ayant de nombreuses caractéristiques en commun (ex. dates de publication, domaines, thèmes, genres, etc.). Le degré de comparabilité d'un corpus varie selon l'ensemble des caractéristiques ou des critères de comparabilité choisis (Rebout, 2012). Dans la figure ci-dessous, nous présentons un schéma de la notion de comparabilité par rapport aux définitions attribuées à ces corpus. Intuitivement, les corpus parallèles peuvent être considérés comme un cas particulier des corpus comparables. Il s'agit de corpus parfaitement comparables.

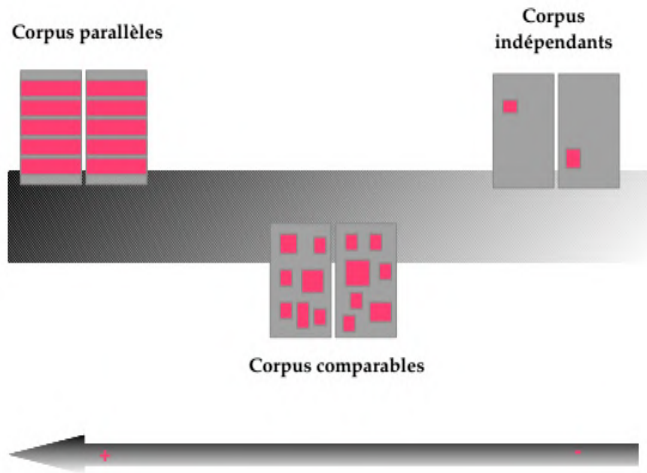


Figure 2. Comparabilité des corpus multilingues.

La particularité des corpus comparables c'est qu'ils ne respectent pas les contraintes imposées par les corpus parallèles. Ils sont largement disponibles, contrairement aux corpus parallèles et les textes qui les composent proviennent généralement de la même source mais ils sont écrits indépendamment dans chaque langue. Ils sont construits à partir de textes originaux plutôt que des textes traduits (corpus parallèles). Ceci permet de réduire le biais de traduction et d'éviter par conséquent l'effet de calque (Rebout, 2012).

La croissance rapide des sources d'informations sur Internet fournit une réelle opportunité pour la construction des corpus comparables. En particulier, les pages de nouvelles issues des agences de presse disponibles en différentes langues, ou encore Wikipédia sont des ressources multilingues volumineuses, riches, exploitables, accessibles et en général libres de droit. Avec l'augmentation des besoins en matière de corpus comparables, la qualité de ces derniers est devenue critique. Le point central de la construction des corpus comparables est l'alignement des documents ou clusters de documents entre langue source et langue cible. Plus les documents alignés qui sont similaires ou comparables, meilleur est l'alignement, et plus le corpus comparable produit est exploitable.

3.3. Corpus comparables versus corpus parallèles

Les corpus parallèles sont une ressource fondamentale utilisée dans la majorité des systèmes de traduction automatique. Généralement, leur structure est basée sur un alignement au niveau de la phrase, qui permet de construire des modèles probabilistes robustes. La difficulté de disposer de cette ressource, surtout dans les domaines spécialisés, a fait émerger les corpus comparables qui sont devenus une alternative aux corpus parallèles. La capacité des corpus comparables à capturer une information plus proche du cadre réel de traduction pouvant éviter les erreurs de traduction littérale, a poussé certains chercheurs à exploiter conjointement les corpus parallèles et comparables pour tirer avantage de ces deux ressources. Ainsi

un troisième axe de recherche fait apparaître plusieurs travaux qui visent à améliorer les systèmes de traduction automatique et l'extraction terminologique bilingue.

Corpus Comparable	Corpus Parallèle
1. Les mots ont plusieurs sens dans le même corpus. 2. De multiples traductions peuvent être associées à un mot. 3. Les traductions pourraient ne pas exister dans le document cible. 4. Les positions et fréquences des mots sont incomparables	1. Un mot n'a qu'un seul sens dans le corpus. 2. Une traduction unique est associée à chaque mot. 3. Il n'y a pas de traductions manquantes entre un corpus source et un corpus cible. 4. Les positions et fréquences des mots en relation de traduction sont comparables.

Tableau 1. Comparatif entre les corpus multilingues

Après l'étude que nous avions menée auprès des travaux destiné à la construction d'un corpus pour une langue peu dotée (Post, et al. 2007), (Gilles, 2002), (Marko, 2000), (Bandmann, et al. 2006), nous avons constaté que malgré la rareté des textes parallèles, et la difficulté de leurs construction, ces corpus permettent d'obtenir des informations sur les caractéristiques des langues, notamment en ce qui concerne la terminologie (Rappazzo, 2014). Les corpus parallèles peuvent paraître plus adaptés à tout type de tâche d'extraction d'informations multilingues puisque l'alignement de phrases y est facilite.

L'intérêt particulier de ce type de corpus, c'est qu'ils peuvent être alignés, le plus souvent, au niveau des phrases, ainsi ils donnent plus de précision pour la transformation linguistique. Egalement, en utilisant les corpus parallèles, nous pouvons consulter leur traduction pour mieux comprendre et découvrir comment une certaine expression du texte de départ a été traduite précédemment et lequel parmi les équivalents possibles a été effectivement utilisé.

Les corpus parallèles constituent donc un élément moteur pour l'extraction des règles syntaxiques, la traduction automatique et la recherche d'information interlingue. Les informaticiens utilisent les ressources obtenues à partir de ces corpus pour améliorer les performances des outils de traduction automatique ou des moteurs de recherche pour le web.

Néanmoins, nous retenons qu'un corpus parallèle nécessite un enrichissement. Afin d'augmenter respectivement la richesse lexicale des catégories existantes, et la variété thématique du corpus, il sera nécessaire d'en ajouter des documents manuellement au fur et à mesure ou d'en extraire à partir des Web (Web-Crawling). Ainsi, Les domaines de recherches actuels se concentrent sur la création automatique de thésaurus multilingue afin d'enrichir les corpus parallèles. Un thésaurus est un type particulier de langage documentaire, il contient un lexique des termes normalisés. Ces termes sont reliés par des relations sémantiques (synonymes, termes associés, termes génériques...) qui illustrent les connaissances du domaine traité. Un thesaurus multilingue est construit en associant à la liste de termes choisis comme descripteurs, c'est-à-dire représentatifs d'un unique concept, tous leurs synonymes dans plusieurs langues (Muller et Langlais, 2011).

4. Alignement des textes parallèles

Selon (Rappazzo, 2014) l'alignement peut être défini comme suit :

Ang : « Aligning consists in finding correspondences, in bilingual parallel corpora, between textual segments that are translation equivalents ».

Fr : « L'alignement consiste à trouver les correspondances entres les segments de deux textes dans un corpus parallel bilingues».

Divers niveaux d'alignement peuvent être définis dans un corpus parallèle. Chaque niveau correspond à un niveau structurel dans le corpus. Par exemple, pour une œuvre littéraire et sa traduction, nous pourrons définir un alignement entre les chapitres, puis entre les paragraphes à l'intérieur des chapitres, puis entre les phrases de ces paragraphes, et même, dans certains cas, au niveau de certains lexèmes à l'intérieur des phrases (Authoul, 2012).

4.1 Alignements par phrases

Nous présentons dans cette section les principales méthodes d'alignement par phrase, en particulier l'ancrage lexical, l'alignement par longueur et les cognats. En 1984, fut la présentation de la première méthode d'alignement automatique nommée : « L'ancrage lexical » il s'agit d'une technique basée sur la distribution des mots, sans l'utilisation d'une source d'information en dehors des deux textes à aligner, c'est-à-dire en observant des cooccurrences de mots à l'intérieur de zones probablement correspondantes. En 1991 fut l'apparition d'une nouvelle méthode d'alignement basé sur la longueur de phrases. C'est une méthode purement statistique, l'idée est que les phrases longues dans le textesource ont tendance à être traduites par des phrases longues dans le texte cible et que les phrases courtes seront traduites par des phrases courtes. Par la suite en 1992, un algorithme qui combine les critères de longueur des phrases à celui d'ancrage lexical a été développé, en utilisant les cognats : des mots apparentésqui ont une origine commune. Le terme peut désigner des mots d'une même langue, ou bien (plus couramment) des mots de langues différentes (Rappazzo, 2014).

4.2 Alignement par mots

Ce type d'alignement reste encore un défi à surmonter. Afin d'obtenir un bon alignement mot à mot, il faut tenir compte des « unités complexes » telles que « mots composés, locutions». Plusieurs méthodes d'alignement de phrases utilisent un alignement de mots. Cependant, dans le cadre de l'alignement de phrases, l'alignement des mots n'est pas le but premier. Lorsque celui-ci est le but premier, les techniques grossières utilisées pour l'alignement des phrases ne sont pas satisfaisantes (Harastani, 2014).

Les mots grammaticaux sont également sources de problèmes : leur correspondance est encore moins nécessaire qu'entre les mots pleins. Néanmoins, il n'est pas possible de les ignorer totalement car ils peuvent faire partie d'une expression à repérer. Les éléments complexes tels que les mots composés ou les locutions, qui sont largement présents dans les phrases, posent également des problèmes cruciaux. Ainsi, par exemple, l'alignement ou l'extraction de lexiques est théoriquement constitué de deux tâches : repérage dans chaque texte et mise en correspondance des termes extraits dans chaque langue. Mais ces tâches

ne peuvent pas être totalement indépendantes car les expressions constituées d'un seul mot graphique dans une langue peuvent être exprimées par plusieurs mots graphiques dans l'autre langue (Miftah, 2016).

4.3 Outils d'alignement

Nous rencontrons aujourd'hui de nombreux logiciels d'alignement automatique, dont nous citons (Miftah, 2016) :

Alinea[3] qui a été développé par Olivier Kraif, directeur du Département d'informatique pédagogique de l'Université Stendhal de Grenoble, est un programme dédié à la constitution et à l'édition de corpus bilingues alignés (Alinéa). Nous trouvons également, AlignFactory[4] qui est un aligneur automatique développé par la société canadienne Terminotix Inc(AlignFactory). L'aligneur au niveau des mots GIZA++[5] estime des modèles de complexité croissante à partir de corpus parallèles (GIZA++). WinAligne[6] est le moins utilisé dans l'alignement de corpus, il fait partie de la suite d'outils de Traduction Assistée par Ordinateur SDL Trados 2007(WinAlign). Cependant, l'outil qui génère un meilleur alignement au niveau de phrases et fournie les statistiques nécessaire sur le corpus aligné est « Alinéa ». Cet outil permet un gain de temps important sur certaines étapes, ainsi qu'une facilité de maintenance légèrement supérieure à celle proposée par les autres outils.

5. Conclusion

Un corpus se caractérise par : une nature (des données langagières), une structure (la sélection, mise en forme et documentation des données) et une finalité (être représentatif d'un phénomène langagier). Un corpus comparable, par opposition à un corpus parallèle, peut être vu comme un ensemble de documents écrits en plusieurs langues traitant des mêmes sujets. Ce type de corpus constitue par son abondance un réservoir d'information utile dans plusieurs domaines d'application tels que les mémoires de traduction ou la catégorisation multilingue.

Dans cet article, nous avons présenté un tour d'horizon sur les corpus en général, plus précisément sur les corpus de textes multilingues parallèles et comparables. Nous avons expliqué en quoi consistent les corpus parallèles et les corpus comparables à travers une comparaison entre les deux types de corpus, nous avons également défini l'alignement, les types, ainsi que des techniques d'alignement qui rendent les textes parallèles utiles et exploitables.

Suie à cette étude, nous avons commencé l'élaboration d'un corpus parallèle aligné au niveau de phrases pour le couple de langues anglais amazighe à l'aide de l'outil Alinéa. Au court terme, nous envisageons enrichir le corpus afin qu'il soit exploitables. Et nous avons également l'intention d'aligner le corpus au niveau de mots.

3 Alinéa: Site personnel de Olivier Kraif :http://olivier.kraif.u-grenoble3.fr/
4 AlignFactory: http://www.terminotix.com/index.asp?name=AlignFactory
5 Giza++:http://www.statmt.org/moses/giza/GIZA++.html
6 WinAlign:http://www.tal.univ-paris3.fr/mkAlign/

Références

Authoul A. (2012). Constitution d'une ressource sémantique arabe à partir de corpus multilingues, Thèse doctorale, Université de Grenoble, Saint-Martin-d'Hères, France

Ataa Allah F. (2015). Traitement automatique des langues peu dotées : Cas de la langue amazighe, Thèse d'Habilitationà Diriger des Recherches, IRCAM, Rabat, Maroc.

Bandmann M. et Sågvall A et Csató J. (2006). Building a Swedish-Turkish Parallel Corpus, Uppsala University, Uppsala, Sweden

Berment V. (2004). Méthode pour informatiser des langues peu dotées, Thèse doctorale, Université Joseph-Fourier, Saint-Martin-d'Hères, France.

Bowker et Pearson (2002). Working with Specialized Language: A practical guide to using corpora. "Working with Specialized Language".

Bouamor D., (2014). Constitution de resources linguistiques multilingues à partir de corpus de texts parallèles et comparables, Thèse doctorale, université parissud, Paris, France.

Gardent C. et Manuélian H., (2005). Création d'un corpus annoté pour le traitement des descriptions définies, L'objet. Volume 8 – n°2/2005.

Guiyao K., (2014). Mesures de comparabilité pour la construction assistée de corpus comparables bilingues thématiques, Thèse doctorale, Université de Bretagne Sud, Lorient, France.

Gilles M. (2002). Web for/as Corpus : A Perspective for the African Languages, Nordic Journal of African Studies 11(2): 266-282.

Hazem A et Emmanuel M., (2013). Extraction de lexiques bilingues à partir de corpus comparables par combinaison de representations contextuelles, Taln-Récital, Les Sables d'Olonne.

Harastani R., (2014). Alignement lexical en corpus comparables : le cas des composés savants et des adjectives relationnels, Thèse doctorale, Université de nantes, Nantes, France.

Harrathi F., (2009). Extraction de concepts et de relations entre concepts à partir des documents multilingues : Approche statistique et ontologique, Thèse doctorale, L'InstitutNationale des Sciences Appliquées de Lyon, France.

Helena M., (2009). Building a Brazilian Portuguese parallel corpus of original and simplified, Advances in Computational Linguistics.Research in Computing Science 41, 2009, pp. 59-70.

Le Serrec A., (2008). Analyse comparative de l'équivalence terminologique en corpus parallèle et en corpus comparable : application au domaine du changement climatique, Thèse doctorale, université de Montréal, Montréal, Canada.

Li et Gaussier (2010). Improving Corpus Comparability for Bilingual Lexicon Extraction from Comparable Corpora, Proceedings of the 23rd International Conference on Computational Linguistics.

Lamraoui F., (2013). Alignement de phrases parallèlesdans des corpus bruités, Thèse doctorale, Université Montréal, Montréal, Canada.

Marko T., (2000). Building the Croatian-English Parallel Corpus, In Proceedings of the Second International Language Resources and Evaluation Conference, 1:523-530

MilošJakubíček, (2013). The TenTen Corpus Family, at 7th International Corpus Linguistics Conference, Lancaster, Royaume-Uni.

Miftah Nina, (2016). Construction de corpus destiné à la traduction automatique des langues peu dotées, Mémoire de projet de fin d'étude, Université Mohammed 5, Rabat, Maroc.

Muller P. et Langlais P. (2011). Comparaison d'une approche miroir et d'une approche distributionnelle pour l'extraction de mots sémantiquement reliés, TALN 2011, Montpellier.

Navlea E. (2014). La traductionautomatiquestatistique factorisée : une application à la paire de languesfrançais-roumain, Thèsedoctorale, université de Strasbourg, Strasbourg, Belgique.

Outahajala M., (2015). Apprentissage d'un étiqueteur morphosyntaxique de la langue amazighe, Thèse doctorale, Ecole Mohammedia d'ingenieurs, Rabat, Maroc.

Pellegrini T., (2008). Transcription automatique de langues peu dotées, Thèse doctorale, Universitéparis-sud, Paris, France.

Post M. et Callison-Burch C. et Osborne M., (2007). Constructing Parallel Corpora for Six Indian Languages via Crowdsourcing, In Proceedings ofWMT'12, pages 401–409, Montreal, Canada.

Rappazzo, Givonna, (2014). Evaluation de la function de recherché dans les outils d'exploitation d'un corpus parallèle, Maîtrise universitaire en traduction, Université de Genève, Suisse.

Rebout L., (2012). L'extraction de phrases en relation de traduction dans Wikipedia, Maîtrise universitaire, Université Montréal, Suisse.

Sellami R. et Sadat F. et Hadrich L., (2013). Traduction Automatique Statistique à partir de corpus comparables : Application au couple de langues arabe-français, CORIA, pages 431–440.

Thi Ngoc Diep Do, (2012). Extraction de corpus parallèle pour la traduction automatique depuis et versune langue peu dotée, Thèse doctorale, Université de Grenoble, Saint-Martin-d'Hères, France.

Yujie Z. et Kiyotaka U. et Qing M. et Hitoshi I., (2004). Building an Annotated Japanese-Chinese Parallel, A Part of NICT Multilingual Corpora, The Tenth Machine Translation Summit.

Zimina-Poirot M., (2004). Approches quatitiatives de l'extraction de resources traditionnelles à partir de corpus parallèle, Thèse doctorale, Université Paris 3, Paris, France.

Analyse textuelle de la ponctuation à l'aide du concordancier ANTONC d'une œuvre de Bélaid At Ali

Noura Tigziri[1] **Ramdane Boukherrouf**[1]
[1] Laboratoire d'aménagement et de l'enseignement de la langue amazighe
Université Mouloud Mammeri de Tizi-Ouzou.
{Nora.tigziri, ramdaneboukherrouf }@gmail.com

Résumé

Notre analyse est une tentative d'application informatique dans le traitement des unités linguistiques. Nous proposons d'utiliser le concordancier *Antconc* pour l'analyse de la ponctuation dans une œuvre de Bélaid At Ali, premier auteur à écrire en kabyle, romans, contes et nouvelles. En effet l'analyse textuelle ou textométrie étudie les textes à partir de formes lexicales, contribue à la caractérisation des genres et des auteurs. On peut l'appliquer à l'analyse du discours ou à la statistique étant donnée qu'elle s'appuie sur des données statistiques et quantitatives.

1. Introduction

Avec les méthodes d'analyse statistique des données textuelles développées depuis les années 1950, nous pouvons mieux décrire et caractériser une langue et le style de documents en tous genres à notre disposition, tels les œuvres littéraires, les discours politiques, etc.

Ces méthodes utilisent des logiciels et axent leurs analyses surtout sur les dénombrements d'occurrences de formes, de lemmes, de catégories grammaticales, etc. dans les textes.

Pour notre part, nous proposons d'utiliser l'un de ces logiciels, à savoir le concordancier *Antconc* pour l'analyse de la ponctuation dans une œuvre de Bélaid At Ali, premier auteur à écrire en kabyle, romans, contes et nouvelles. En effet l'analyse textuelle ou textométrie étudie les textes à partir de formes lexicales, contribue à la caractérisation des genres et des auteurs. On peut l'appliquer à l'analyse du discours ou à la statistique étant donnée qu'elle s'appuie sur des données statistiques et quantitatives.

2. Ponctuation et prosodie

La non ponctuation ou la mauvaise ponctuation d'un texte provoquent un manque de lisibilité certain par la difficulté et l'ambiguité rencontrés dans leur lecture. Ceci n'étant guère étonnant quand on sait le rôle joué par la prosodie dans l'organisation syntaxique

de l'amazigh. La ponctuation transpose dans l'écrit les faits de la langue orale. Grevisse considère la ponctuation

comme étant un ensemble de signes conventionnels servant à indiquer, dans l'écrit, des faits de la langue orale comme les pauses et l'intonation, ou à marquer certaines coupures et certains liens logiques. [...]. (1980 : 144)

La ponctuation contient un certain nombre de signes typographiques conventionnels qui sont: le point(.), le point d'interrogation (?), le point d'exclamation (!), la virgule (,), le point-virgule (;), les deux points (:), les points de suspension (...), les crochets [], les parenthèses (), les guillemets « », le tiret (-) et la barre oblique(/).

La prosodie (pause) joue un rôle important dans la segmentation phrastique et les différentes unités textuelles. En effet, la pause joue un rôle syntaxique (subordination et coordination, séparation de l'indicateur du thème du syntagme prédicatif, le nom apposé, etc.) et la fonction énonciative (hésitation, non-dit, etc.)

La pratique de la ponctuation dans le domaine berbère, s'est développée avec les productions écrites depuis le XVIIIème siècle avec la publication de plusieurs œuvres (Bensedira 1887, Boulifa 1913, Bélaid Ait Ali 1950, etc.). A Partir des années 1970, la production écrite kabyle a pris un essor considérable avec la production des œuvres littéraires (romans, nouvelles, etc.). Cependant, cette multitude de productions a engendré des divergences énormes en matière de codification de la ponctuation.

Au niveau académique, mis à part quelques réflexions superficielles (Chaker 2009, Mahrazi 2014, Boukherrouf 2014, Guerchouh 2014, Amaoui 2014), la codification de la ponctuation du berbère n'a fait l'objet d'aucune étude approfondie consacrée exclusivement à la proposition d'un système de codification à utiliser dans les productions écrites. L'intégration du berbère dans les nouveaux domine d'usage, notamment dans l'enseignement et la communication, obligent les chercheurs à prendre la question comme préoccupation incontournable des recherches actuelles du domaine berbère. Cependant, le traitement de la ponctuation ne peut être pris en charge indépendamment des autres domaines linguistiques : prosodie, syntaxe, sémantique, stylistique et texte.

Dans le cadre de cette recherche, ce qui nous intéresse le plus sont les points de suspension. Le point de suspension est apparu au XVIIe siècle. Il a été dénommé par plusieurs appellations tels – le point interrompu, les points multiples, les points de coupure, les points poursuivant, les points suspensifs, mais d'après Alain Rey, le nom « points de suspension » est attesté en 1752.

D'une manière générale, les points de suspension expriment l'inaccompli, le non-dit ou une idée inachevée. Ainsi quand ils apparaissent en fin de phrase, on pourrait imaginer soit que l'auteur fait une pause qui peut exprimer plusieurs hypothèses sur le narrateur : une hésitation, un silence, une réflexion qui se prolonge. Ce caractère de ponctuation sert à exprimer plus que ce qu'on veut dire.

Nous avons choisi de travailler sur le roman *Lwali Bu udrar* de Bélaid At Ali[1], premier auteur d'expression kabyle. Pour pouvoir l'utiliser dans AntConc, nous utilisons sa version électronique. Notre corpus est un texte intégral choisi pour un objectif précis : nous avons remarqué la multiplicité des points de suspension dans ce texte et nous analysons les raisons et les contextes de leur apparition. En effet comme le souligne Sinclair, 1996, «A corpus is a collection of pieces of language that are selected and ordered according to explicit linguistic criteria in order to be used as a sample of the language» (Julien Rault 2015).

3. Présentation de l'outil d'analyse : Le logiciel Antconc

Nous avons choisi AntConc qui est un logiciel d'analyse textuelle développé par Laurence Anthony, professeur à l'université de Waseda au Japon[2].

2.1. La page d'accueil

Le logiciel *AntConc* comporte un certain nombre de fonctionnalités apparentes à droite et en haut de la page d'accueil (Cf. Figure 1). Parmi ces différentes fonctionnalités qu'il propose, nous n'en avons retenu que quelques-unes, à la fois facilement accessibles et pertinentes pour notre projet. Il s'agit de la « Word List » ou liste de mots où on retrouve toutes les unités lexicales avec le nombre de leur occurrence, utilisées dans le texte, de « concordance» qui nous donne le contexte d'apparition de la forme lexicale , à partir d'une occurrence, au contexte large (phrase ou paragraphe ; texte souvent, accessible grâce à « l'ascenseur »), de « concordance plot » qui nous éclaire sur la distribution d'une forme lexicale dans le corpus.

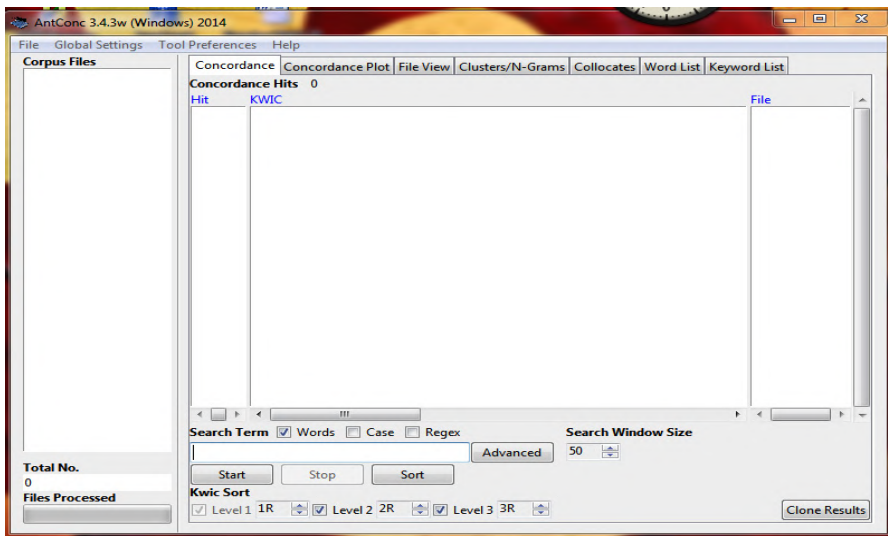


Figure 1. Page d'accueil du logiciel AntConc

1 Ecrivain kabyle, né en 1909, décédé en 1950. Son œuvre est publié en deux volumes, intitulés « les écrits de Bélaid Ait Ali» en 1964 par J-M. Dallet et J-L. Degezelle.

2 Téléchargeable sur : http://www.antlab.sci.waseda.ac.jp/antconc_index.html

2.2. Contraintes techniques

L'utilisation d'*AntConc* nécessite un certain nombre de réglages et de contraintes. Pour le corpus, il doit répondre à ces exigences :

- Utilisation d'une police UNICODE avec un codage UTF8.
- Le format de fichier à utiliser est le txt ou .html

Au niveau de « global settings », on procéder à ce qui suit :

- Réglage des éléments graphiques non-séparateurs de mots (ajout des graphèmes spécifiques à Tamazight pour que le logiciel reconnaisse les textes écrits en tamazight)

3. Présentation et Analyse du corpus

3.1. Présentation du corpus

Dans ce récit de Lwali Bu drar- qui raconte l'histoire d'un homme, simple d'esprit, sujet à des moqueries dans son village, devenu un imam dans un autre village par pur hasard et à cause d'une rivalité qui opposait deux personnes de ce village pour cette place d'imam tant convoité-, nous avons une utilisation excessive des points de suspension comme nous l'avons signalé précédemment. Les contextes d'emploi de ces points de suspension seront décrits ci-dessous.

3.2. Analyse des données

Sur 33546 occurrences de syntagmes, presque la moitié des formes spécifiques sont différentes (8359). 5910 hapax à savoir les mots que l'auteur utilise une seule fois sont assez importants par rapport aux nombres de formes totales (8559). La « world List » donne les résultats comme suit (Cf. figure 2) :

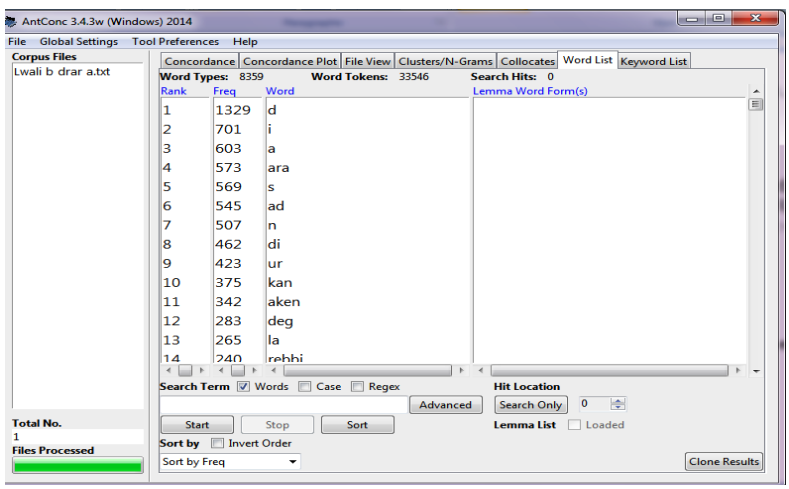


Figure 2. World List

3.3. Recherche des points de suspension

La recherche montre que l'auteur a utilisé 227 fois les points de suspension dans son texte ce qui représente un chiffre considérable. Il viennent au 15ème rang ce qui place ce signe de ponctuation parmi les signes les plus utilisés dans ce corpus, ce que montre l'importance qu'accorde l'auteur aux points de suspension (figure 3).

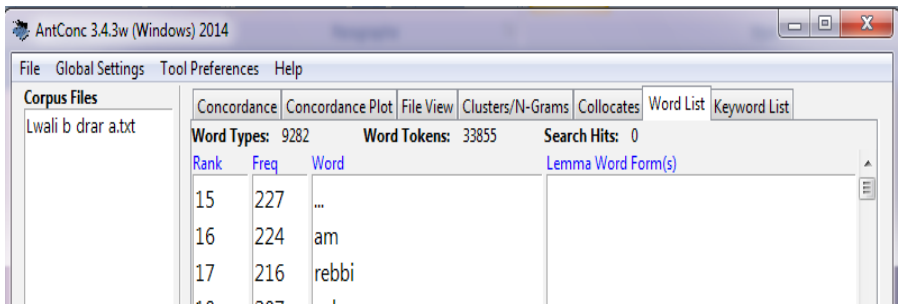


Figure 3. Recherche des points de suspension

La deuxième fonctionnalité « concordance plot » montre la distribution des points de suspension dans *Lwali Bu Udrar* (figure 4).

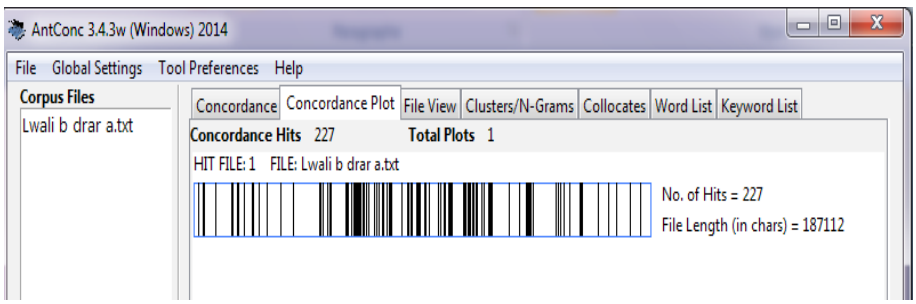


Figure 4. concordance plot de la distribution des points de suspension dans Lwali Bu Udrar

Cette distribution montre que les points de suspension se retrouvent du début jusqu'à la fin du texte avec cependant une concentration au milieu. En cliquant sur cette concentration, le logiciel nous renvoie au texte (figure 5) :

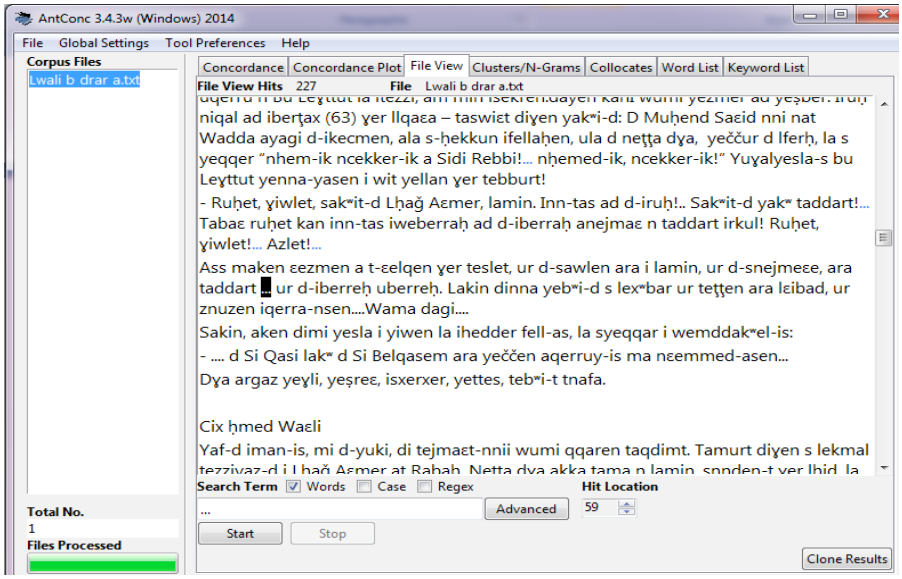


Figure 5. Distribution des points de suspension dans Lwali Bu Udrar

3.4. Analyse de l'apparition des points de suspension

Même si les points de suspensions comme sigle de ponctuation occupent des contextes syntaxiques, leur fonction est d'ordre énonciatif. En effet, l'ensemble des contextes d'utilisation des points de suspension employé par l'auteur dans notre corpus relève d'une utilisation individuelle, et qui véhiculent des fonctions exclusivement énonciatives. Dans notre travail, nous avons relevé deux contextes d'utilisation des points de suspensions : intraphrastique et à la fin de la phrase. Notre analyse dégage les fonctions suivantes :

3.4.1. L'hésitation et le changement d'avis

L'interruption d'une phrase : La phrase ne se termine pas en raison d'une hésitation ou d'un changement d'avis : *a k-d-iniɣ…ruḥ, ruḥ, dayen* ou l'hésitation à l'intérieur d'une phrase : *waqila…ihakka i d-yenna* !

3.4.2. L'énumération

Dans ce cas l'auteur fait une énumération écourtée à la fin de sa phrase : *truḥ ar tmurt n l'Espagne, n legliz, n França…*

3.4.3. Le tabou

Pour des raisons de contraintes sociales ; les tabous sociaux, l'auteur substitue l'unité lexical par les points de suspension. *D cix kan u... hadi-k makan*

3.4.4. L'hésitation, le doute

Dans le cas où l'auteur n'est pas certain de son discours, il fait recours aux points de suspension: *iɛedda limin-nninaγ yeḥnet… mbɛed. Lḥaṣun jerbaγ nek.Daγen ţatura… tura….*

Nekini?... Isem-iw Bu...bu...bu...

Wala-k sgellin…i… la tesmeḥsiseḍ itxuniyin tketbeḍ. Aɛni … dayen ţdekkirent,

Wissen… Lḥaṣun, ayen ur yezmir ara bnadem a t-id-yinnid agi daγa dayen

Nek daγa γur-iaqerru… icfawat… dayenkan.

Lexmisb^{w}edraragi… daγa zdeffir wedrar

Sakin… sakin, dayent abaɛ isi beddaγ, dehcaγ

Lḥaṣun… Lḥaṣun daγa s γur tilwin-nni I d-sliγ kra imeslayen aken

ladγa, d waken ara tfetamkan ilaqen, yerna waqila,... teţuγam cwit ulaγar yedrimen

Wala-k sgellin…i… la tesmeḥsiseḍ itxuniyin tketbeḍ. Aɛni … dayen

3.4.5. La complicité avec l'interlocuteur

Les points de suspension peuvent avoir une fonction argumentative. En effet, le locuteur cherche la complicité de l'interlocuteur en employant des points de suspension : *ma dayen iɛeddan, daγa aken qqaren waɛraban:lifat mat…*

Dayen ţdekkirent, nnaγ?...

Ula d lɛarc-nni… wissen a Rebbi.

Awufan am ass-agi...

A tebra tin akenar ayaru Rebbi deg qarruy-iw urak-ţ-mliγ tamsalt alama tceḍhaḍ-ţ!…

3.4.6. L'étonnement

Comme utilisation individuelle des points de suspension, nous avons relevé dans notre texte, la fonction d'étonnement employée par l'auteur.

Sddaw daγa talemast b^{w}edrar, almiiyi-ḍhar … a Rebbi! Ma d sbɛa leswaq igɛamren

Sakin… sakin, dayen tabaɛ isibeddaγ, dehcaγ.

Ur t-sineḍ ara Cix ḥmedWaɛli ??!… Ad aγ-d-yenfaɛRebbi s lbaraka-s?!Ur t-sineḍ ara taddart inna iwumi qqarenTizi-n-Tfilkut?…

3.4.7. Le non-dit

L'un des liages de connexion textuelle des discours et le non-dit. Dans notre corpus, nous avons enregistré la fonction implicite des points de suspension est assez répondante.

Lḥaṣun jerbaγnek.Daγen ţatura… tura…

Tura mi ɛeddan wussan

Daken llan lhan
Awufan am ass-agi...
Dayneţţa tura acḥal d abridi… rriγ lxir deg arrac
Yiwen d sin lekdub atγarwiyi ḍyak^{w}... zayed naqes.
yerna ad ak-dsenɛat allen-nsent ma… tizegzawinneγtiber kanin
lliγimi-w... mbla ma nniγ-ak.
wanda… iyi-čebbi Rebbi ɛusen-iyi-d yergazen
yeznuzuy-asentiţemceqlal yid sent γef summa, aɛni… marcinwarnaγamek
TideţiRebbi… yefraḥcwitnnif-iw. A bnadem… tamurt-ik t tamurt-ik.
Isem-is tideţur tɛuhdeγaraurtemliγ… imi d nek s yiman-iwurtesineγara ass-a.
ţdekkirent, nnaγ?…
Ula d nekhuzzaγkanaqerruy-iw, nniγ-as :
- Ayih
'riγ-t di syenakindayenuriyi-ţensar ara. Yenna-k :
- Ihi s taɛrabt ?
Nniγ-as :
- Ala.
Yenna-k :
Ihi ... s rumit?
Nniγ-as:
- Ala.
Lakin, segmitfehmeγicuk aɛniţkellixaγ yes, stebɛaγ-as:
- S teqbaylit.

4. Conclusion

Cette modeste analyse, peut être considérée comme un travail préliminaire à une étude plus approfondie sur l'analyse automatique des contextes et des différentes fonctions des sigles de ponctuations employés par les textes écrits kabyles.

Au terme de cette étude, en comptabilisant tous les points de suspension dans tous les contextes d'apparition nous avons constaté ceci :

151 points de suspension ont été utilisés pour le non-dit. Belaid ait Ali laisse le lecteur deviner ce qu'il ne veut pas dire explicitement. Il fait appel à l'imagination du lecteur. Très peu pour éviter un tabou. Nous l'avons retrouvé une seule fois. 25 points de suspension ont été employés par l'auteur pour marquer son hésitation lors de la narration d'un évènement. On a l'impression qu'il n'est pas très sûr de ce qui s'est passé exactement.18 points de suspension sont employés pour marquer l'étonnement de l'auteur. 6 points de suspension marquent l'hésitation, le doute dans le texte 7 points de suspension met en relief plus la complicité avec le lecteur

Comme on le voit, la majorité des points de suspension trouvés dans le texte (151) se réfèrent au non-dit, à l'implicite. Si on regarde avec attention ce récit, il n'est point étonnant de trouver ce résultat étant donné que l'objet essentiel décrit par Belaid At Ali est les croyances populaires de l'époque, en majorité païennes qui côtoient les religions monothéistes. Ce récit écrit pendant l'occupation française explique l'état d'ignorance dans lequel était plongé la population et comment ces croyances et religions partageaient l'imagination des gens. En plaçant les points de suspension dans certains passages, Belaid At Ali ne voulait pas brusquer les croyances des gens même si on sentait que lui n'y croyait pas !

De même Belaid At Ali, posait le problème de l'identité à travers la langue kabyle. Il se demandait lui, imam, ne connaissant pas l'arabe comment il allait s'adresser à Dieu qui lui, ne connaissait pas le kabyle mais il finit par comprendre qu'il peut s'adresser à Dieu dans la langue qu'il veut. Ce questionnement de l'identité revient principalement dans ce passage où en entendant tuxuniyin (qui vient de lexwan) psalmodier, il se demandait dans quelle langue elles parlaient :

- Ihi s taɛrabt ?

Nniɣ-as :

- Ala.

Yenna-k :

Ihi ... s trumit?

Nniɣ-as:

- Ala.

Lakin, segmitfehmeɣicukaɛniṭkellixaɣyes, stebɛaɣ-as:

- S teqbaylit.

Références

Duez D.(1991). La *pause de la parole de l'homme politique,* Paris, Edition du CNRS, collection Son et Parole.

Boukherrouf R. (2006). *Structures intonatives des énoncés verbaux complexes en berbère(kabyle) coordination et subordination,* Mémoire de Magistère, Université Mouloud Mammeri de Tizi-Ouzou.

Boukherrouf R. (2013). «Le rôle de la sémantique et de la prosodie dans la distinction entre la subordination et la coordination sans marque monématique de jonction», *Faits de syntaxe amazighe,* Actes du colloque internantional organisé par le Centre d'Aménagement Linguistique, IRCAM, Rabat Maroc, pp.55-65.

Chaker S (1991). «Eléments de prosodie berbère : quelques données exploratoires», in EDB N°8, PP.5-25.

Chaker S. (2009) : « Structuration prosodique et structuration (typo-) graphique en berbère: exemples kabyles.», in *Etudes de phonétique et linguistique berbère, Hommage à Naima Louali* (1961-2005), PEETERS, Paris Louvain, PP.227-242.

Challah S. (2004). *Le rôle de l'intonation entre en syntaxe. Etude de cas portant sur l'opposition de l'état (analyse intonosyntaxique de quelques types d'énoncés)*, Mémoire de Magistère, Université Mouloud Mammeri de Tizi-Ouzou.

Grevisse M. (1980). *Le bon usage- grammaire française avec des remarques sur la langue française d 'Ajourd 'hui*, Duculot.

Mettouchi, A. (2003). « Focalisation contrastive et structure de l'information en *kabyle (berbère)», Mémoires de la Société de Linguistique* de Paris, PP. 179-196.

Mahrazi M. (2014). Le passage de l'oralité à l'écriture de l'amazighe : problème de la ponctuation in *Actes du 2ème colloque international* « la langue amazighe, de la tradition orale au champ de la production écrite », Bouira, 2014

Amaoui Mahmoud (2014). Quelques éléments de réflexion pour servir à la codification de la ponctuation du berbère, in *StudiAfricanistici* 3, Langues et Littératures Berbères : Développement et Standardisation, Napoli 2014.

Tigziri Noura (2016). Analyse de l'album ISEFRA Lounis Ait Menguellet à l'aide de ANTCONC in colloque international « *Litt'Orale* », De la littérature orale comme patrimoine, Regards croisés, les 7 et 8 avril 2016, Oujda (à paraitre)

Tigziri N. (2016). Analyse textuelle à l'aide du concordancier ANTONC d'une œuvre de BélaidAt Ali (Bu yedmimen et tassa), Colloque international «BELAID AIT ALI (1909-1950). *Un auteur et une œuvre à (re) lire*», les 24-25 avril 2016, Tizi-Ouzou (à paraître)

Tigzri N. (2014). La réalisation de grands corpus berbères normalisés et interopérables : enjeu culturel et enjeu d'ingénierie linguistique in Revue *ASINAG N 9*, Rabat, Maroc.

Convertisseur numérique : Tifinaghe - Braille

Nouria Yakoubi[1] Jamal Frain[2] Fadoua Ataa Allah[2]
[1] Association Charte pour le développement social
nourianouria.yakoubi@gmail.com
[2] Centre des Etudes Informatiques, Systèmes d'Information et de Communication, IRCAM
{frain, ataaallah}@ircam.ma

Résumé

Dans la perspective de la généralisation de l'enseignement et de l'apprentissage de la langue amazighe au profit de toutes catégories de la société marocaine, ce travail propose un système d'écriture de la graphie tifinaghe à base de braille. Ce système assura aux non-voyants d'apprendre la lecture et l'écriture, et de les mettre sur le même pied d'égalité que les voyants dans la réussite scolaire en amazighe.

Le système, entretenu dans cet article, distingue entre deux ensembles de l'alphabet tifinaghe. Les caractères du premier ensemble sont représentés par les points braille à la base d'une règle, proposée par Mme Nouria Yakoubi, qui aide les gens ayant une déficience visuelle à reconnaître et se rappeler facilement de ces caractères écrits en braille. Tandis que les caractères du deuxième ensemble sont extraits de la langue anglaise, arabe et française, selon leur correspondance phonétique.

Cet article introduit également une application de conversion, dotée d'un clavier amazighe virtuel parlant pour aider les personnes aveugles de s'assurer de ce qu'elles écrivent et d'éviter les erreurs. Elle comporte un module de conversion, permettant de substituer, automatiquement, chaque caractère tifinaghe par son correspondant dans le système braille, et vice versa. En outre, cette application favorise la sauvegarde des textes écrits en tifinaghe ou en braille dans des fichiers de type Word. De plus, elle permet de convertir les textes des fichiers Word écrits en tifinaghe à braille et vice versa directement à la pression du bouton de conversion, tout en respectant la mise en forme, assurant ainsi aux aveugles d'imprimer les textes en braille.

1. Introduction

L'intégration sociale et l'accès à l'éducation des personnes handicapées en général, et les non-voyants en particulier, représentent un défi civilisationnel pour toute communauté, notamment dans les pays sous-développés ou en voie de développement. Cependant, cette question, principalement, d'ordre moral et humanitaire, nous interpelle à donner plus d'importance aux handicapés visuels, afin d'améliorer leurs conditions de vie.

D'ailleurs, cette catégorie, qui représente une partie non négligeable, est engagée au service de notre communauté. Par sa capacité à la participation et à la créativité, elle joue un rôle

considérable dans l'enrichissement de l'héritage national, via plusieurs domaines, et en particulier le domaine de la science, tels que la justice et la littérature. D'où la nécessité de lui assurer les moyens fondamentaux pour bénéficier aussi de l'enseignement et l'apprentissage de la langue amazighe.

Dans cette perspective, un projet d'écriture de la langue amazighe en braille a été proposé, notamment dans l'absence d'un tel projet assurant aux non-voyants la lecture et l'écriture du tifinaghe en braille, que ça soit par les méthodes traditionnel via une tablette, ou par les moyens technologiques récents tels que l'ordinateur, les téléphones intelligents et les tablettes tactiles.

Ce projet vise, dans un premier lieu, la réalisation d'un système d'écriture de l'alphabet tifinaghe en braille, afin d'assurer aux personnes ayant une déficience visuelle un moyen d'apprendre à lire et à écrire en amazighe. Et dans un deuxième temps, il s'intéresse à l'élaboration d'un convertisseur permettant la substitution des caractères tifinaghes en braille et vice versa, pour faciliter l'échange entre les personnes qui n'ont pas d'handicap visuel et les non-voyants.

2. L'écriture amazighe en braille

S'attacher à la langue amazighe, et en faire un axe de recherche et d'étude, nous suscite à s'intéresser à l'écriture amazighe, qui constitue en soi un support de conservation et de transmission de l'héritage culturel marocain. D'ailleurs, le système d'écriture amazighe préconisé par l'Institut Royal de la Culture Amazighe (IRCAM) ''le tifinaghe'' n'est qu'un répertoire de signes multimillénaire conservés dans les motifs décoratifs de l'artisanat nord-africain (Ameur et al., 2004). C'est un système qui trace l'histoire des amazighes et valorise leurs pensées et sentiments dans des écrits qui méritent d'être lus et étudiés par toutes personnes y compris les non-voyants.

Dans ce contexte, nous proposons un système de transcription en braille des 33 caractères tifinaghes utilisés dans le système éducatif marocain. Cette méthode tactile à points saillants, élaborée par le Français «Louis Braille» en 1829, a rendu possible l'apprentissage de l'écriture et de la lecture chez les aveugles et les malvoyants ; ce qui leur a permis d'accéder à l'information et de s'informer sur leur environnement.

La méthode braille repose sur six points saillants, représentés sous la forme d'une matrice à six (6) points, composée de trois (3) lignes et de deux (2) colonnes. Ces points permettant aux personnes avec déficience visuelle de lire avec leurs mains, du bout des doigts. Ainsi, le système braille est constitué d'un ensemble de cellules, dont chacune, représentant un caractère, est formée par un (1) à six (6) points en relief. Ces points sont conventionnellement numérotés de haut en bas et de gauche à droit (*cf.* Figure 1) (UNESCO, 1949).

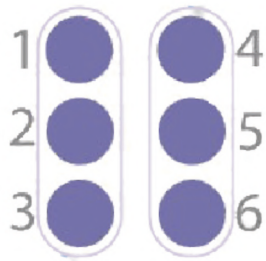


Figure 1: Cellule braille

Etant l'objet de notre projet est d'élaborer un système d'écriture en braille pour la langue amazighe, nous avons proposé un signe pour distinguer le tifinaghe du reste des graphies transcrites par la méthode braille. Ce signe est illustré par la Figure 2, ci-dessous :

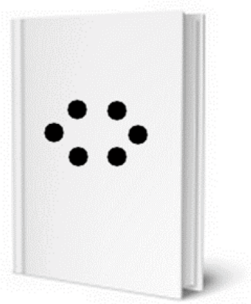


Figure 2 : Signe de la graphie tifinaghe en braille

Le système de transcription en braille proposé pour la graphie tifinaghe, se base sur la conjointure d'un ensemble de règles, cités ci-dessous, afin de simplifier l'apprentissage de la lecture et l'écriture de la langue amazighe aux non-voyants et les aider à se rappeler de l'ensemble des caractères :

- La fréquence d'usage de l'ensemble des caractères, illustrée par la figure 3 (Ataa Allah, 2015), a été étudié dans la représentation des caractères, d'une manière à que les caractères les plus utilisés dans la langue amazighe sont représenté en braille avec le minimum de points, afin d'optimiser le temps et l'effort d'écriture. Ainsi, la majorité des caractères les plus fréquemment utilisés sont représentés par deux(2), trois (3) ou quatre (4) points au maximum. Cependant, le reste des caractères est représenté par cinq (5) points ; à l'exception du caractère '⵰' [a] et du caractère '⵿' [e], qui sont représentés respectivement par un point et six (6) points, vue que le premier est le caractère le plus utilisé en amazighe et le deuxième par contre est le moins utilisé.

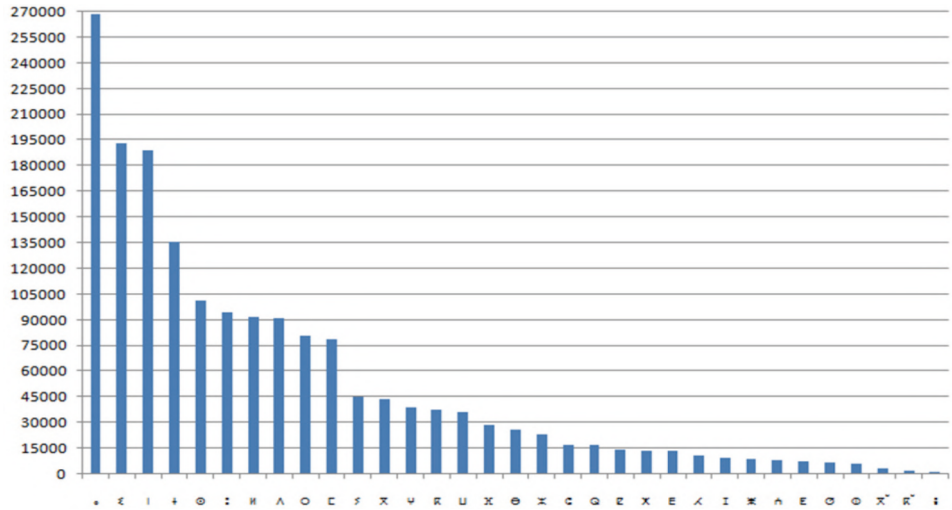


Figure 3 : Occurrences des caractères tifinaghes

- Le respect des représentations braille désignées à la ponctuation, d'une manière à ne pas les affecter à lettre.
- La relation phonétique de l'emphase entre l'ensemble des lettres ('ⴷ' [d], 'ⵔ' [r], 'ⵙ' [s], 'ⵜ'[t], 'ⵣ'[z]) et ('ⴹ' [ḍ], 'ⵕ' [ṛ], 'ⵚ' [ṣ], 'ⵟ' [ṭ], 'ⵥ' [ẓ]) et celle de la labiovélarisation marquée par l'ajout de la lettre modificative 'ⵯ' [w] auⵅ caractères 'ⵅ' [g] et 'ⵇ' [k] sont marquées en braille par l'ajout du point 6 sur la deuxième colonne de la cellule de chaque lettre, comme expliqué par le schéma de la figure 4.

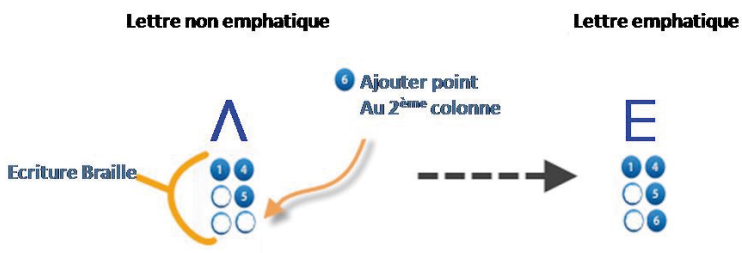


Figure 4 : Règles de la représentation des lettres emphatiques et labiovélaires

- La correspondance phonétique entre des lettres de la langue amazighe et celles des langues anglaise, arabe ou française, qui a été exploitée pour simplifier le processus de l'écriture et de la lecture de l'amazighe pour ceux qui maîtrisent ces langues. A titre d'exemple, la figure 5 illustre le cas de la lettre 'ب' en arabe, 'B' en français et

'ⵖ' en amazighe qui se prononcent de la même manière, d'où l'intérêt qu'ils soient transcrites par la même lettre braille, représentée par les points 1 et 2.

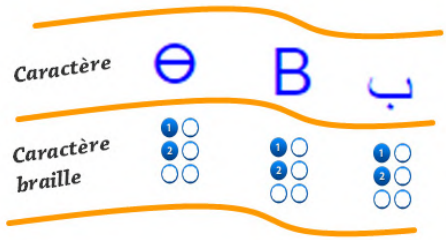


Figure 5 : La représentation de la lettre 'ⵖ' [b]

Le tableau 1 présente la liste des lettres tifinaghes dont la prononciation et la représentation en braille sont similaires à la fois à des lettres en arabe et en français. Cependant, le tableau 2 regroupe la liste des lettres extraites de l'arabe qui ont la même prononciation. Similairement, tableaux 3 et 4 présentent respectivement la lettre extraite de la langue française et anglaise.

#	Caractère tifinagh	Appellation en français	Caractère braille
1	ⴰ	ya	
2	ⵖ	yab	
5	ⵃ	yad	
8	ⵄ	yef	
9	ⵅ	yak	
11	ⵀ	yah	
17	ⵉ	yaj	

#	Caractère tifinagh	Appellation en français	Caractère braille
18	ⵍ	yal	
19	ⵎ	yam	
20	ⵏ	yan	
22	ⵓ	yar	
25	ⵙ	yas	
28	ⵜ	yat	

Tableau 1 : Liste des lettres extraites de la langue arabe et française

#	Caractère tifinagh	Appellation en français	Caractère braille	#	Caractère tifinagh	Appellation en français	Caractère braille
7	ⵢ	yey	⠽	24	ⵖ	yagh	⠣
12	ⵅ	yahh	⠭	27	ⵛ	yach	⠩
14	ⵅ	yakh	⠭	29	ⵟ	yatt	⠾
15	ⵇ	yaq	⠿	30	ⵡ	yaw	⠺
21	ⵓ	you	⠥	31	ⵣ	yay	⠵

Tableau 2 : Liste des lettres extraites de la langue arabe

#	Caractère tifinagh	Appellation en français	Caractère braille
3	ⵖ	yag	⠣

Tableau 3 : Lettre extraite de la langue française

#	Caractère tifinagh	Appellation en français	Caractère braille
16	ⵄ	yi	⠔

Tableau 4 : Lettre extraite de la langue anglaise

En outre des lettres traitées précédemment, nous signalons que le choix de la représentation de la lettre ‘ⵓ’ [u] par les points 2 et 6 est réalisée à la base, d’une part, de sa similarité en sa prononciation avec la diacritique ‘ⵓ’ [u] de la langue arabe, et d’autre part, l’intérêt d’utiliser un nombre réduit de points, notamment que cette lettre est classée 6ème dans la liste des lettres les plus en usage en amazighe. Cependant, il a été opté que la représentation de la lettre française ‘z’ soit affectée à la lettre ‘ⵣ’ [ẓ] au lieu de ‘ⵥ’ [z], qui ont la même prononciation, et ceux pour respecter la règle de l’emphase, vue que le point n° 6 fait partie des points représentant le ‘z’.

Eu égard des quatre règles expliquées ci-dessous et de ce qui précède, le tableau 5 ci-dessous présente l'ensemble des lettres brailles proposées pour représenter les 33 caractères de la graphie tifinaghe.

3. Description du convertisseur

Il est bien constaté plus que jamais que la technologie bouscule les habitudes, la tradition et la manière dont nous accédons au Savoir. Elle permet de rendre les choses possibles et plus faciles. Néanmoins, ce n'est pas évident que toute personne a le pouvoir de profiter des avantages de la technologie, particulièrement les personnes handicapées visuelles. Dans la perspective de remédier à cette déficience, il est important de réaliser des outils d'appui au profit des non-voyants, leur permettant d'améliorer leurs conditions de vie, notamment les conditions de leur apprentissage.

#	Caractère tifinagh	Appellation en français	Caractère braille	#	Caractère tifinagh	Appellation en français	Caractère braille	#	Caractère tifinagh	Appellation en français	Caractère braille
1	ⵄ	ya		12	ⵅ	yahh		23	ⵕ	yarr	
2	ⴱ	yab		13	yaa	ⵀ		24	ⵖ	yagh	
3	ⴳ	yag		14	ⴾ	yakh		25	ⵙ	yas	
4	ⴳⵯ	yagw		15	ⵇ	yaq		26	ⵚ	yass	
5	ⴷ	yad		16	ⴹ	yi		27	ⵛ	yach	
6	ⴹ	yadd		17	ⵊ	yaj		28	ⵜ	yat	
7	ⵉ	yey		18	ⵍ	yal		29	ⵟ	yatt	
8	ⵃ	yef		19	ⵎ	yam		30	ⵡ	yaw	
9	ⴽ	yak		20	ⵏ	yan		31	ⵢ	yay	
10	ⴽⵯ	yakw		21	ⵓ	you		32	ⵣ	yaz	
11	ⵀ	yah		22	ⵔ	yar		33	ⵥ	yazz	

Tableau 5 : Liste des représentations brailles de l'ensemble des lettres tifinaghes

Dans ce contexte, nous avons proposé de réaliser une application de convertisseur automatique de tifinaghe vers braille et vice versa, dotée d'une voix humaine. Cette application servira, aux non-voyants, comme outil d'appui pour assurer leur indépendance dans le processus de la lecture et de l'écriture en langue amazighe. A cette fin, l'application proposée est composée de :

- une interface interactive, dotée d'une voix humaine, permettant de guider l'usager dans l'utilisation et l'exploitation de l'application. Elle contient un menu qui permet

le choix du système d'écriture (tifinaghe ou braille), une zone de saisie et une autre d'affichage.

- un clavier amazighe virtuel parlant, permet à la fois la saisie en braille et en tifinaghe selon le choix de l'utilisateur, et l'aide à s'assurer de ce qu'il écrit pour éviter les erreurs.
- un module de conversion instantanée, permettant de substituer chaque caractère saisi par son correspondant, de copier et de sauvegarder les textes résultants dans des fichiers de type Word.
- un module de conversion de fichiers, qui assure la conversion des textes saisis dans des fichiers Word écrits en tifinaghe à braille et vice versa automatiquement à la sélection du fichier à convertir, tout en respectant la mise en forme[1], assurant ainsi aux non-voyants la possibilité d'imprimer les textes en braille.

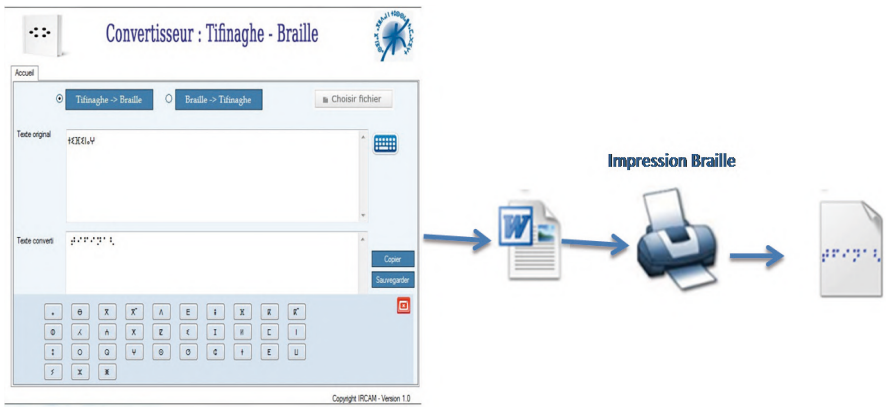


Figure 3 : Processus de conversion et d'impression en braille

4. Conclusion

Dans la perspective d'encourager et d'appuyer l'apprentissage de la lecture et de l'écriture de la langue amazighe transcrite en tifinaghe, au profit des personnes ayant une déficience visuelle, le présent article propose un système d'écriture en braille et présente un convertisseur automatique de braille vers tifinaghe et vice-versa. Ce dernier permettra d'ouvrir de nouveaux horizons pour les non-voyants et les aidera à répondre à un ensemble de besoins vitaux, notamment l'élaboration d'une base de données lexicales en langue amazighe et la publication des articles et des manuels scolaires en langue amazighe.

1 Pour plus d'information sur la méthodologie adoptée pour la préservation de la mise en page, veuillez consulter (Ataa Allah et Frain, 2013).

Références

M. Ameur, A. Bouhjar, F. Boukhris, A. Boukouss, A. Boumalk, M. Elmedlaoui, E. M. Iazzi, H. Souifi, (2004). Initiation à la langue amazighe. Rabat, Maroc : IRCAM.

F. Ataa Allah, J. Frain, (2013). Amazigh Converter based on WordprocessingML, Actes de la 6ème édition de Language & Technology Conference (LTC 2013), Poznań, Pologne, 7-9 décembre 2013.

F. Ataa Allah (2015). Traitement automatique des langues peu dotées : Cas de la langue amazighe. Mémoire d'Habilitation à Diriger des Recherches, à l'Institut Royal de la Culture Amazighe.

UNESCO (1949). Rapport sur la situation mondiale du système Braille, L'éducation, la science et la culture.

TICAM'16, p. 109 - 114

Arabicized and Romanized Berber Automatic Identification

Wafia Adouane[1] **Nasredine Semmar**[2] **Richard Johansson**[3]
[1] Department of FLoV, University of Gothenburg, Sweden
wafia.gu@gmail.com
[2] CEA Saclay – Nano-INNOV, Institut CARNOT CEA LIST nasredine.semmar@cea.fr
[3] Department of CSE, University of Gothenburg, Sweden
richard.johansson@gu.se

Abstract

We present an automatic language identification tool for both Arabicized Berber (Berber written in the Arabic script) and Romanized Berber (Berber written in the Latin script). The focus is on short texts (social media content). We use supervised machine learning method with character and word-based n-gram models as features. We also describe the corpora used in this paper. For both Arabicized and Romanized Berber, character-based 5-grams score the best giving an F-score of 99.50%.

1. Introduction

Automatic language identification (ALI) is the first crucial step to automatically analyze and process any natural language. It is the identification of the natural language of an input text/speech by a machine. Obviously, a wrong identification of the language at hand outputs misleading information. ALI is a well-established task since early 1960's where various methods have been applied to a wide range of general purpose languages. Berber or Tamazight is an under-resourced language and unknown to the state-of-the-art automatic language identifiers. Arabicized Berber (AB) or Berber written in the Arabic script is misclassified as Arabic (precisely as Modern Standard Arabic)[1] and Romanized Berber (RB) or Berber written in the Latin script is confused with unrelated languages. Berber has its own unique script called Tifinagh which is hardly supported by the available technology devices. It does not have one standardized writing system, so each country uses its own orthography. For convenience, Berber is also written in Latin and Arabic[2] scripts in its formal form.

To the best of our knowledge, only some works have been done to automatically recognize Berber, for instance (Chelali *et al.*, 2015) created a Berber speaker identification system using some speech signal information as features. Also (Halimouche *et al.*, 2014) used prosodic

1 Among the freely available language identification tools, we tried Google Translator, Open Xerox language and Translated labs at http://labs.translated.net.
2 Using Arabic script to transliterate Berber has existed since the beginning of the Islamic Era (L. Souag, 2004).

information to discriminate between affirmative and interrogative sentences in Berber. Both works were done at the speaker level. There are other Natural Language Processing (NLP) applications for Berber but always assuming that the input is in Berber.

In this paper, we first describe how we built the linguistic resources (corpora) used to build our automatic language identifier. We then describe the used methods and experiments. We conclude with the general findings and some future directions.

2. Linguistic resources

The approach we use in this paper, namely supervised machine learning requires training dataset and since AB and RB are under-resourced languages, we built our own dataset for both. We collected content from the Web where we counted each user's comment as a single document. In total, we collected 7,000 documents for RB mostly from Berber websites promoting Berber culture and language (4,900 documents), YouTube (910 documents), news websites (700 documents) and Facebook (490 documents). The document's average length is 30 tokens.

Likewise for AB, we collected 503 documents, mainly content from forums (234 documents), blogs (160 documents) and Facebook (109 documents). For more data, we have selected varied topics from Algerian newspapers and segment them. Originally the news texts are short, 1,500 words each. So we have considered each paragraph as a document (maximum of 178 words). Then, we have added 1,497 documents to the ones collected from social media to get 2,000 documents in total. We could have collected content only from social media platforms, but we wanted to include as many Berber varieties as possible and the selected newspapers use different Berber varieties spoken in Algeria. Hence, we made sure to include content from the six most popular Berber varieties, namely Kabyle, Tachelhit, Tachenwit, Tachawit, Tamzabit and Tarifit.

With the help of a Berber speaker from Morocco (precisely from Taza), we checked that all the included documents are in one of the Berber varieties. The task is easy 'is a document written in Berber or something else' compared to 'in which Berber variety each document is written in'. It is hard for a Berber native speaker not to distinguish Berber from other languages like Arabic or French. Therefore, we assume the inter-annotator agreement to be satisfactory. For historical reasons, Berber uses lots of mixed-languages, in particular words borrowed from Arabic and French depending on the Berber variety. Consequently, we allowed mix-language documents given that they contain clearly Berber words (in Arabic or Latin script) and a native speaker can understand or produce the same (sounds very natural for a native speaker).

Arabicized Berber (AB) does not use special characters and it coexists with Maghrebi Arabic where the dialectal contact has made it hard for non Maghrebi people to distinguish it from local Arabic dialects [3]. For instance each word in the Arabicized Berber (Kabyle) sentence أحمل ساقول ماشي دول كان which means [love is from heart and not just a word] has

3 In all polls about the hardest dialect to learn, Arabic speakers mention Maghrebi Arabic which has Berber, French and words of unknown origins which is not the case of other Arabic dialects.

a false friend in Arabic. In Modern Standard Arabic (MSA), the sentence means literally [I carry I will say going countries was] which is none sense. It has also different meanings in different Arabic dialects. This motivates us to build a system which is able to detect Berber and distinguish it from Arabic. Therefore, we added a dataset (built for an ongoing project) containing 16000 documents written in the eight most popular Arabic varieties, namely Algerian (ALG), Egyptian (EGY), Gulf (GUL), Levantine (LEV), Mesopotamian (KUI), Moroccan (MOR), Tunisian (TUN) dialects and MSA.

For Romanized Berber (RB), the main issue is the use of non-standardized orthography. We simply included different spellings found in the collected data. RB could be easily confused with some unknown languages to the available automatic language identifiers such as Romanized Arabic (RA), called also Arabizi, which is a close language to Maltese and which has lots of false friends with Romanized Persian. Therefore, we added a dataset of 3,000 documents including 500 documents for each of English (EN), French (FR), Maltese (ML), Romanized Arabic (RA) and Romanized Persian (RP). The added languages are likely to be confused with RB, particularly short texts, and we want to avoid that. For all languages, we introduced some normalization where we reduced all the adjacent repeated letters to maximum two characters, removed unimportant tokens such as punctuation, emoticons, stop-words and named entities (organization, product, location and people's names). The reason is that we want to build a system which learns linguistic information rather than learning topical or country specific words.

3. Methods and experiments

We use Support Vector Machines method as implemented in Scikit-learn package (F. Pedregosa *et al.*, 2011) [4]. As features, we use both character-based and word based n-gram of different lengths. We use the term frequency-inverse document frequency[5] (TF-IFD) scheme to filter the features based on their importance. In all experiments, we use the binary classification setting as opposed to 9-class classification for AB or 6-class classification for RB. We conduct separately two experiments based on the used script, one for languages using Arabic script (the eight Arabic varieties and AB) and the other for languages using Latin script (EN, FR, ML, RA, RB and RP). For the 1st group, we use a balanced dataset of 18 000 documents (2 000 for each mentioned language) where for each we use 80% for training and the rest for evaluation. For the 2nd group, we use a dataset of 3 000 documents (500 for each of the mentioned languages) where 300 documents for each language are used for training and 200 for evaluation.

Arabicized Berber

From Scikit-learn package, we have experimented with different methods and parameters and found that the LinearSVC classifier (we refer to it as SVM for simplification) using the default parameters outperformed all the rest settings. So, we use it in these experiments. We experiment with different character and word n-gram lengths with the dataset using the

4 For more information see: http://scikit-learn.org/stable/

5 A statistical measure used to filter stop-words and keep only important words for each document.

Arabic script. The SVM performance is shown in Table 1.

The results show that increasing the n-gram length increases the classification accuracy where the character 5-grams performs the best with an accuracy of 89.63%. Also, combining unigram with trigrams performs slightly better compared to using only unigrams or trigrams, accuracies of 87.22%, 55.44 % and 87.11% respectively. However, the classifier's performance using word-based n-gram decreases with increasing the n-gram length. Word unigram outperforms other n-grams with an overall accuracy of 87.83%.

	Accuracy (%)	
Features	Character-based	Word-based
unigram	55.44	87.83
bigrams	78.05	51.36
trigrams	87.11	19.30
-4grams	89.02	13.72
-5grams	89.63	12.80
-2+1grams	77.97	87.58
-3+1grams	87.22	87.66

Table 1: SVM performance with different n-gram lengths

In Table 2, we show the performance of the SVM classifier per language using character-based 5-grams.

Language	**Precision (%)**	**Recall (%)**	**F-score(%)**
AB	99.75	99.25	99.50
ALG	87.10	81.00	83.94
EGY	94.80	82.00	87.94
GUL	84.25	88.25	86.20
KUI	90.09	95.50	92.72
LEV	92.24	77.25	84.08
MOR	89.64	93.00	91.29
MSA	86.06	98.75	91.97
TUN	84.95	91.75	88.22

Table 2. SVM performance per language

The SVM classifier accurately detects Arabicized Berber, an F-score of 99.50% and easily distinguish it from Arabic varieties.

Romanized Berber

We redo the same experiments with the dataset using the Latin script. Table 3 shows the performance of the SVM using different n-gram lengths. The same as with the 1st experiment, increasing the length of the character-based n-gram has a positive effect on the classification, 5-grams scores the best 98.91% accuracy. Combining character unigram with trigrams has slightly improved the classification, 98.66% accuracy compared to using only unigram 95.58% or trigrams 98.58%. Nevertheless, increasing the length of the word-based n-gram has caused a decrease in the classification accuracy, unigram outperforms the rest of n-grams with an accuracy of 96.16%. Combining word unigram with bigrams has improved the classifier's performance 96.33% accuracy.

	Accuracy (%)	
Features	**Character-based**	**Word-based**
unigram	95.58	96.16
bigrams	98.41	79.41
trigrams	98.58	58.66
4-grams	98.58	43.08
5-grams	98.91	30.33
1+2-grams	97.99	96.33
1+3-grams	98.66	95.91

Table 3: SVM performance with different n-gram lengths

For illustration, we show in Table 4 the performance of the SVM classifier per language using character-based 5-grams as features.

Again, the classifier accurately detects RB. This might be related to the use of some special characters which are not used in other languages.

Language	**Precision (%)**	**Recall (%)**	**F-score(%)**
EN	98.51	99.00	98.75
FR	99.00	99.50	99.25
ML	99.50	99.00	99.25
RA	98.98	97.50	98.24
RB	99.00	100	99.50
RP	97.52	9850	98.01

Table 4: SVM performance per language

4. Conclusion

We have presented an automatic language identification tool for Berber written in both Arabic and Latin scripts. SVM classifier using character-based 5-grams performs the best. It accurately distinguishes Arabicized and Romanized Berber from other close languages with an F-score of 99.50%. Word-based unigram performs also reasonably well. We can conclude that machine supervised machine learning is a well suited method for automatic identification of Arabicized and Romanized Berber. We assume that the system will work well for longer documents since it performs well for shorter documents. As future work, we want to distinguish between Berber varieties and build linguistic resources for each to be able to automatically analyze and process them.

References

Chelali, F. Z., Sadeddine, K. and Djeradi, A. (2015). Speaker identification system using LPC-application on Berber language. HDSKD journal, Vol. 01, No. 02, pp. 29-46.

Halimouche, R., Teffahi, H. and Falek, L. (2014). Detecting sentences types in Berber language. International Conference on Multimedia Computing and Systems (ICMCS), pp. 197- 200.

Souag, L. (2004).Writing Berber languages: a quick summary. Archived from http://goo.gl/ooA4uZ, Retrieved on May 26th, 2016.

Dunning, T. (1994). Statistical Identification of Language. Technical Report MCCS, pp. 94-273, Computing Research Lab (CRL), New Mexico State University.

Pedregosa, F., Varoquaux, G., Gramfort, G., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M. & Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12, pp. 2825–2830.

Muntsa, P. and Llus, P. (2004). Comparing methods for language identification. Procesamiento del Lenguaje Natural, 33, pp. 155–162.

L'analyse morphologique de la langue amazighe: état de l'art et perspectives

Rachid Ammari[1] Lahbib Zenkouar[1] Mohammed Outahajala[2]
[1] Ecole Mohammedia d'Ingénieurs
ammari.rachid@hotmail.fr, lzenkouar@yahoo.fr
[2] Institut Royal de la Culture Amazighe
outahajala@ircam.ma

Résumé

L'intégration des Technologies de l'Information et de Communication (TIC) à l'apprentissage de la langue amazighe est nécessaire pour un meilleur accompagnement de la veille technologique. A l'instar de la majorité des langues peu dotée, la langue amazighe connaît une pénurie en terme de recherche dans tous les niveaux en occurrence, l'analyse morphologique. Dans cette optique, ce travail présente l'état de l'art de l'informatisation de l'amazighe, et plus précisément l'analyse morphologique, les travaux effectués basés sur le modèle des états finis, les deux plateformes NOOJ ou SFST, et les perspectives à entreprendre afin de promouvoir une telle démarche.

1. Introduction

La langue amazighe du Maroc est une composante essentielle de la culture marocaine grâce à sa large utilisation et son usage important dans la vie quotidienne. Selon le dernier recensement de 2014, 27% de la population marocaine l'utilise sous forme de plusieurs dialectes. Cependant, et au fil des années la langue amazighe n'a cessé de provoquer une grande polémique au sein de la société marocaine, chose qu'a donnée naissance à la création d'une nouvelle institution gouvernementale qui est l'Institut Royal de la Culture Amazighe (IRCAM). Ce point d'inflexion a permis à la langue amazighe de garantir une présence dans les différents domaines actifs à savoir l'enseignement, l'administration et le secteur médiatique. Cet acquis a permis à la langue amazighe de se doter d'une graphie officielle, d'un codage propre dans le standard Unicode ainsi que des normes appropriées pour la disposition d'un clavier amazighe. En outre, une démarche de construction des lexiques (Kamel, 2006; Ameur *et al.*, 2009) de mise en place des règles de segmentation de la chaîne parlée ainsi que l'élaboration des règles grammaticales (Ameur *et al.*, 2006), (Boukhris et. al., 2008) a été réalisée.

2. Historique de la langue amazighe

La langue amazighe ou Tamazight (ⵜⴰⵎⴰⵣⵉⵖⵜ) fait partie de la famille des langues afro-asiatiques ou chamito-sémitiques (Greenberg, 1966; Ouakrim, 1995 ; Chaker, 1991). Elle

présente la langue d'une population appelée «Imazighen» qui se présente à l'heure actuelle dans une dizaine de pays allant du Maroc jusqu'à l'oasis égyptienne de Siwa, en passant par l'Algérie avec 25%, la Tunisie, la Mauritanie, la Libye le Niger et le Mali (Chaker, 2003). Au Maroc, l'amazighe se répartit selon trois grandes zones régionales : le Tarifit au Nord, le Tamazight au Maroc central et au Sud-Est et le Tashelhit au Sud-Ouest et dans le Haut-Atlas. Chacun de ses dialectes comprend des sous dialectes ou dialectes locaux. Depuis la création de l'IRCAM, il s'est engagé à réaliser un processus de standardisation (Ameur *et al.*, 2004a), dans le but d'uniformiser les structures et d'atténuer les divergences, en éliminant les occurrences non distinctives qui entraînent souvent des problèmes d'intercompréhension.

2.1. Alphabet amazighe

Comme toutes les langues qui passent de la voie orale au mode écrit, la langue amazighe se trouvait dans l'obligation d'avoir un système graphique. Au Maroc, le choix s'est orienté sur Tifinaghe pour des raisons techniques, historiques et symboliques. Par conséquent, l'IRCAM a développé un alphabet basé sur un système graphique ayant une tendance phonologique. Ce système ne conserve pas toutes les réalisations phonétiques produites, mais seulement ceux qui sont fonctionnelles. Il est composé de 27 consonnes, 2 semi-consonnes, 3 voyelles pleines et une voyelle neutre.

2.2. Encodage Unicode

Une fois le système Tifinaghe est adopté comme graphie officielle au Maroc pour la langue amazighe, l'encodage Tifinaghe est devenu nécessaire. Dans ce sens, des travaux énormes ont été déclenchés par le centre des études et systèmes d'information et de communication de l'IRCAM, ce qui a donné naissance à un codage Unicode qui s'articule autour de quatre sous-ensembles de caractères Tifinaghe, dont nous citons :

- L'ensemble de base de l'IRCAM,
- L'ensemble étendu de l'IRCAM,
- Autres lettres néo-Tifinaghe
- Lettres Touareg moderne.

2.3. La morphologie amazighe

L'amazighe est une langue morphologiquement riche et compliquée. Du point de vue morphosyntaxique, la langue amazighe contient : noms, verbes, pronoms et des mots de fonction qui comprennent à leur tour les adverbes, prépositions, etc. On citera dans la partie suivante une description microscopique de la morphologie amazighe.

2.3.1. Les caractéristiques des noms

Le nom dans la langue amazighe est composé d'un mot entre deux espaces et ne commence pas par un article (Oulhaj, 2000). Le nom dans la langue amazighe est divisé en trois catégories noms propres, noms communs et noms dérivés.

• Nom Propre

Il s'agit de nom d'une personne (ⵎⵓⵀⴰ [moha] « Moha ») ou de lieu ⴰⵣⵔⵓ [azru] « pierre ».

• Le nom commun

Le nom commun est divisé en deux types abstrait et concret. Celui-ci peut être soit animé (ⵜⴰⵎⵜⵓⴷⵜ [tamtudt] « femme ») ou inanimé (ⵜⵉⴳⵎⵎⵉ [tigmmi] « maison »).

• Le nom dérivé :

Le nom dérivé est issu d'une racine verbale. Il est composé d'une préfixation ou suffixation d'un morphème de dérivation plus des variations intra-radicales.Il est caractérisé par le genre (masculin ou féminin), le nombre (singulier ou pluriel), et l'état (libre / construit).

• Genre

Les noms masculins commencent par 'a' , 'i', ou 'u' avec quelques exceptions, qui incluent les noms de parenté. Pour transformer un nom masculin singulier au féminin, le morphème 't' est ajouté au début et à la fin. Exemple : Ifri [tifrit] « petite grotte ».

• Nombre

La plupart des noms singuliers masculins dans la langue amazighe commencent par 'a', les noms qui commencent par un «i» sont moins fréquent, suivi par les noms commençant par 'u'. La forme plurielle est dérivée du singulier avec l'addition de quelques affixes et certains changements vocaliques. Ce dernier affecte les lettres initiales dans le nom: le son 'a' passe à ' i' ou à 'u' (Sadiqi, 1997) . Pour marquer le pluriel, le son «an» est ajouté à la fin pour un nom masculin, occasionnant un changement de voyelle si le nom singulier se termine par une voyelle. Pour la plus part des noms féminins, le «a» après l'initial 't ' change à 'i', le «t» final est remplacé par «in».

Exemple :

- asrdun ..... (mule, m. sg.) ===isrdunan ..... (mules, m.pl.)
- ifri .... (grotte, m.sg.) ===ifran ..... (grottes, m.pl.)
- uccen3 ..... (loup, m.sg.) ===uccenan ..... (loups, m.pl.)
- tamUart ..... (femme, f. sg.) ===timUarin ..... (femme, f. pl.)

Le pluriel s'articule autour de quatre types : le pluriel externe, interne, mixte et le pluriel en ⵉⵏ [id] (Boukhris *et al.*, 2008)..

Type de pluriel	Obtention du pluriel avec exemple
Externe	alternance vocalique accompagnée par une suffixation de ⵏ [n] ou l'une de ses variantes (ⴹⵏ [in], ⵄⵏ [an], ⵄⵖⵏ [ayn], ⵍⵍ [wn], ⵄⵍⵍ[awn], ⵍⵄⵏ [wan], ⵍⴹⵏ [win], ⵜⵏ [tn], ⵖⴹⵏ [yin]). ⵄⵅⵅⵄⵣ [axxam]-> ⴹⵅⵅⵄⵣⵏ [ixxamn] "maisons", ⵜⵄⵔⵄⵜ [tarbat] "fille" -> ⵜⴹⵔⵄⵜⴹⵛⵜ [tirbatin] "filles".
Interne	une alternance vocalique plus un changement de voyelles internes. ⵄⴰⵔⵄⵔ [adrar] -> ⴹⴰⵙⵄⵔⵄⵔ [idurar] "montagnes".
Mixte	alternance d'une voyelle interne et/ou d'une consonne plus une suffixation par ⴹ [n]. ⴹⵍⴹ [ili] "part"-> ⴹⵍⵄⵏ [ilan] "parts", ou bien par une alternance vocalique initiale accompagnée d'un changement vocalique final ⵄ [a] plus une alternance interne ⵄⵎⵅⵅⵄⵔⵙ [amggaru] "dernier" > ⴹⵎⵅⵅⵙⵔⵄ [imggura] "derniers".
en ⴹⴰ [id]	une préfixation ⴹⴰ [id] du nom au singulier

Tableau 1. type de pluriel avec exemple

• ***Etat***

Les noms sont connus par deux états: l'état libre et l'état construit. L'état libre est la forme normale de tous les noms. Chaque état, à son tour, est divisé en deux catégories : abstrait ou concret. Un nom est à l'état construit quand le changement touche la première voyelle et ce dans les contextes suivants (Sadiqi et Ennaji, 2004) :

- Lorsque le nom suit le verbe dans une phrase et il est le sujet du verbe ;
- Lorsque le nom suit une préposition ;

Dans l'état construit, les noms féminins enlèvent le «a» après l'initial «t», mais dans les noms masculins :

- Initial 'a' change à 'u'
- Initial 'i' change à 'y'
- Initial 'u' change à 'wu' (Sadiqi, 1997).

2.3.2. Les caractéristiques des verbes

Les verbes sont des mots graphiques simples avec leurs inflexions (personne, nombre, aspect) ou morphèmes dérivatives (Boukhris *et al.*, 2008)..

2.3.3. Les pronoms

Les pronoms dans la langue amazighe sont soit démonstratifs, exclamatifs, indéfinis, interrogatifs, personnels, possessifs ou relatifs. Un exemple de leurs utilisations peut être

comme suit:

- Pronom personnel (ⵏⴽⴽ [nkk] « moi »)
- Pronom possessif (ⵡⵉⵏⵓ [winu] « le mien »);
- Pronom démonstratif (ⵡⴰⴷ [wad] « ce »);
- Pronom interrogatif (ⴽⵔⴰ[kra] « quelque chose »)

2.3.4. Les adverbes

Les adverbes sont classés en fonction de leur sémantique, on trouve les adverbes de lieu, de temps, de quantité, de manière et les adverbes interrogatifs.

Exemple : [qqaH] « tous », comme dans l'exemple: ⵇⵇⴰⵄ ⵜⵉⵏⵏⵉ ⵏⴰⵜⵜⵉⵏⵉ [qqaH middn da ttinin] « tous les gens parlent »

2.3.5. Les focalisateurs

Les interjections et les conjonctions sont écrites en un mot entre deux espaces ou bien un espace et un point de ponctuation. Par exemple : ⵜⵔ [mr] « si ». /

2.3.6. Les prépositions

Les prépositions sont indépendantes par rapport au nom qu'elles précèdent. Cependant, si la préposition est suivie d'un pronom personnel, celle ci forme avec le pronom personnel une seule chaîne délimitée par des blancs ou bien un blanc et une marque de ponctuation. Par exemple : ⵖⵔ [Gr] « à, vers » + le pronom ⵉ [i] « moi » donnent ⵖⴰⵔⵉ/ⵖⵓⵔⵉ [(Gari/ Guri)] « à moi, avec moi »; qui sont des variantes de la préposition ⵖⵔ [Gr] quand elle est utilisée conjointement avec le pronom personnel ⵉ [i].

2.3.7. Les particules

Les particules sont souvent isolées. Elles sont de plusieurs types:

- Les particules aspectuelles telles que ⴰⵇⵇⴰ [aqqa], ⴰⵔ [ar], ⴰⴷ [ad];
- La particule de négation ⵓⵔ [ur];
- Les particules d'orientation, comme ⵏⵏ [nn]« là-bas » ;
- La particule de prédication ⴷ [d].

2.3.8. Les déterminants

Les déterminants prennent toujours la forme d'un seul mot délimité par deux espaces. Ils sont divisés en articles, démonstratifs, exclamatifs, articles indéfinis, interrogatifs, chiffres ordinaux, possessifs, présentatifs et quantificateurs. ⴽⵓⵍⵍⵓ [kullu] « chaque »est un exemple de quantificateur.

2.3.9. Les marques de ponctuation

Les marques de ponctuation en amazighe marocain sont similaires aux marques de ponctuation adoptées par les langues internationales. Elles ont les mêmes fonctions.

3. Analyse morphologique de la langue amazighe

L'analyse morphologique est l'une des tâches importantes en PNL et qui présente une base pour plusieurs applications. C'est une étape fondamentale des divers niveaux des applications à savoir la traduction automatique, la génération automatique de textes et la correction orthographique.

3.1. Etat de l'art sur les approches computationnelles

Au cours des 25 dernières années, le Finite State Technology (FST) ou bien la technologie des états finis a eu un grand impact sur une variété des applications du Traitement Automatique des Langues (TAL) et l'ingénierie linguistique grâce à la puissance des automates des états finis.

3.1.1. Two-LevelMorphology

Développé par Koskenniemi en 1980 cette solution dépend fortement des méthodes à états finis. Il s'agit du premier modèle dans l'histoire de la linguistique computationnelle pouvant analyser et traiter les modèles morphologiques les plus complexes. Un des plus grands avantages de cette solution est que toutes les règles sont déclaratives : en effet, la formulation originale permet à la fois l'analyse et la génération au sein de la même grammaire.

3.1.2. XFST

Le Xerox Finite State Tool est l'un des outils les plus sophistiqués pour la construction et le traitement des applications basés sur le modèle d'états finis. Le XFST fournit une boîte à outils pour un traitement linguistique plus puissant.

3.1.3. SFST

Le Stuttgart Finite State Transducer est un système open-source, librement disponible. Il fournit un ensemble d'outils basés sur le modèle des états finis permettant l'analyse, la génération et le traitement des automates et transducteurs. Ce système tourne sous linux.

Bien que ces technologies à états finis présentent un certain nombre d'avantages, ils présentent également plusieurs désavantages. L'handicap majeur est qu'elles fournissent un seul formalisme censé, normalement, être utilisé pour décrire tous les phénomènes linguistiques. Contrairement à la technique précédente, l'environnement de développement linguistique NooJ ou SFST offre plusieurs outils formels destinés à faciliter la description de chaque phénomène linguistique.

3.2. Travaux et évaluation sur la langue amazighe

3.2.1. Sur La plateforme NooJ

NooJ, publié en 2002 par Max Silberztein (Silberztein, 2007), est une plate-forme permettant de développer les exigences linguistiques, de construire et de gérer des dictionnaires électroniques et de grammaires formelles à large couverture. Le but est de formaliser les différents phénomènes linguistiques à savoir l'orthographe, la morphologie flexionnelle et dérivationnelle, le lexique (de mots simples, mots composés et expressions figées), la syntaxe (locale, structurelle et transformationnelle), la désambigüisation, la sémantique et les ontologies. Une fois lesdites descriptions sont formalisées à l'aide des machines à états finis, elles deviennent applicables au traitement des textes et corpus de taille importante. Le module morphologique de NooJ, se base sur les expressions régulières pour effectuer des recherches et des traitements dans les textes tout en associant leurs reconnaissances à la liste d'annotations correspondante (étiquette grammaticale, des informations sémantiques, la traduction, etc.). Pour mettre en place ce concept, des techniques ont été mises en place appelées opérateurs de transformations prédéfinis dans NooJ, à savoir :

- LW : positionner le curseur au début de mot
- RW : positionner le curseur à la fin de mot
- R : déplacement vers la droite.
- L : déplacement vers la gauche
- B : effacer le dernier caractère.
- S : effacer le caractère courant.

Dans la suite de ce travail nous allons citer les travaux déjà réalisés dans la cadre de l'analyse morphologique basée essentiellement sur la combinaison de plusieurs transducteurs des états finis tout en utilisant l'environnement NooJ ou SFST.

• Construction du lexique amazighe

La première étape qui précède chaque analyse morphologique d'une langue est le développement d'un lexique morphologique. Pour se faire, une collecte manuelle des données a été faite concernant les listes des noms, des verbes et des pronoms. A l'aide des techniques de Nooj, bon nombre des règles a été défini, et ce, selon les types de morphologie (dérivationnelle ou flexionnelle).Trois lexiques amazighes (Nejme *et al.,* 2016) ont été créés à savoir :

- « NAmLex» , signifie « Lexique des noms amazighes», contient une liste des noms simples, composés et dérivés
- « VAmLex» , signifie « Lexique des verbes amazighes», contient une liste des verbes simples et dérivés.
- « PAmLex» , signifie « Lexique des particules Amazighe», contient une liste des pronoms.

• ***Système d'analyse automatique de la morphologie nominale***

Il s'agit de la construction d'un système d'analyse automatique de la morphologie nominale de l'amazighe standard du Maroc tout en profitant des avantages des modèles à états finis au sein de l'environnement linguistique de développement NooJ et en faisant appel à des règles grammaticales à large couverture (Nejme *et al.*, 2016).

Ce travail commence par l'analyse lexicale avec la formalisation flexionnelle et dérivationnelle du lexique des noms. En outre, la formalisation de la classe des noms propres se base sur le dictionnaire électronique « EDicAMPN » (ElectronicDictionary for Amazigh Proper Noun) contenant 424 entrées de prénoms simples et proprement amazighes et 500 entrées de prénoms étrangers, sous format simple et composé.

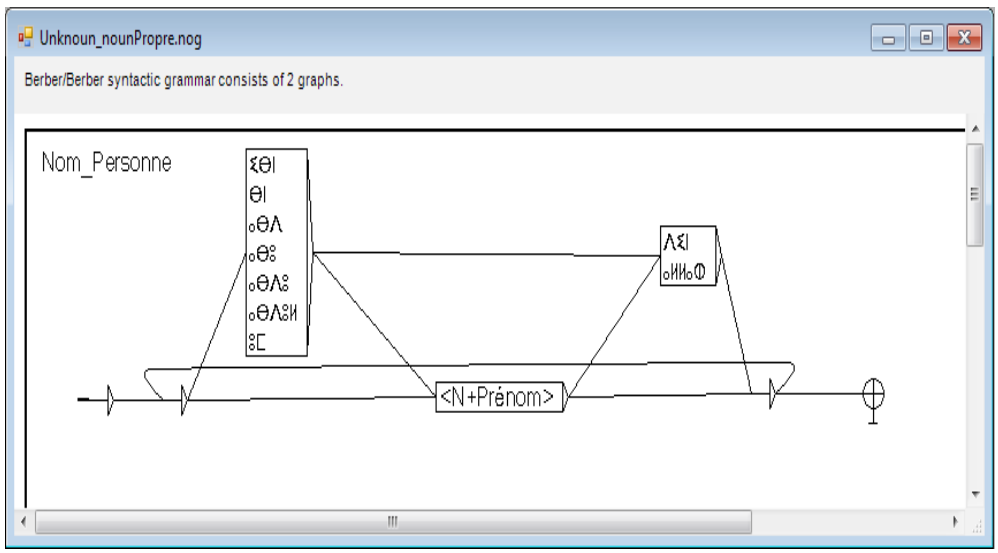


Figure 1 .Exemple d'un graphe montrant la grammaire morphologique pour la reconnaissance des prénoms composés sur NOOJ.

La règle dans NooJ	Explication	Exemple
<LW>+<RW>+	La règle ajoute un + [t] au début et à la fin du nom	ⵙⵓⵍⵉ [isli] "marié" -> +ⵙⵓⵍⵉ+[tislit] "mariée".

Figure 2 . Exemple de règles grammaticales permettant de passer d'un nom masculin à un nom féminin

Test et évaluation

Pour l'évaluation des ressources linguistiques, un corpus a été construit comprenant :

- Une liste contenant 4524 entrées de noms communs simples et fléchies,
- Une liste basée sur la nouvelle grammaire de l'amazighe contenant 113 entrées de noms dérivés,
- Une liste de 200 entrées de prénoms étrangers, sous format simple et composé, transcrits.

		Nombre de formes reconnues		**Nombre de formes non reconnues**	
		Nombre	%	Nombre	%
Résultats	Noms communs	4205	93%	319	7,05%
	Noms propres	594	95,19%	28	4,5%
	Noms dérivés	103	91,15%	10	8,84%

Tableau 2 – Evaluation des ressources linguistiques

Cette analyse a réalisé un bon chiffre en terme de couverture d'environ 93% et 7,42 % des mots non traités. Les écarts sont dus principalement aux problèmes de flexion et de transcription.

3.2.2. Sur SFST

• The Stuttgart Finite State Transducer – SFST

C'est une plateforme open source qui permet le traitement, la génération des ressources et l'implémentation des analyseurs morphologiques. Ce système contient les modules suivants (Schmid, 2007) :

- FSTS - PL , un langage de programmation pour la mise en œuvre des transducteurs à états finis,
- Un compilateur pour TSN – compilateur,
- Outils d'application, la comparaison et l'impression FST,
- Bibliothèque FST.

• ANMorph: développement et Evaluation

Le travail de Raiss (Raiss et Sforza, 2011) présente un analyseur / générateur morphologique à deux niveaux pour les noms amazighs, basé sur les techniques de SFST. Il a été développé en utilisant le LCTL. ANMorph contient 15 règles, 1541 noms et seulement 9.32% qui ne sont pas analysés. Cette situation est due au fait que certains mots échappaient aux règles prédéfinies (Raiss et Sforza, 2011).

4. Conclusion et perspectives

Nous avons abordé dans ce travail la problématique de l'analyse morphologique. Une des difficultés principales de cette tâche était de formaliser la morphologie flexionnelle ainsi que dérivationnelle pour les catégories grammaticales nom (nom simple et dérivé) et verbe. Pour pallier à ce problème un système est construit regroupant un ensemble de grammaire flexionnelle et dérivationnelle ainsi qu'une grammaire pour la reconnaissance des noms et verbes en se basant sur la technologie à états finis incorporé dans la plateforme de développement linguistique NooJ ou SFST. Après traitement, les résultats obtenus sont satisfaisants et encourageants.

Dans le futur proche nous envisageons :

- Terminer la collecte d'un corpus amazighe, riche de point de vue qualitatif et quantitatif avec moins de mots hors vocabulaire,
- Faire une étude comparative des résultats obtenus avec ceux cités dans ce travail,
- Améliorer le travail existant en ajoutant d'autre règles afin d'inclure tous les cas exceptionnels pour la langue amazighe.

Références

Ameur, M., Bouhjar, A., Boukhris, F., Boukouss, A., Boumalk , A., Elmedlaoui, A., Iazzi, M., Souifi, H. (2010). Introduction to the Amazigh language. Publications de Institut Royal de la Culture Amazighe.

Ameur, M., Bouhjar A., Boukhris F., Boukouss A., Boumalk A., Elmedlaoui M. and Iazzi E. (2006a). Graphie et orthographe de l'Amazighe. Publications de Institut Royal de la Culture Amazighe.

Ameur, M., Bouhjar A., ElMedlaoui M., and M. Iazzi (2006b).Vocabulaire de la langue amazighe (Français – Amazighe). Publications de Institut Royal de la Culture Amazighe.

Ataa Allah, F. and Boulaknadel S. (2010). Light morphology processing for Amazighe Language. In Proceedings of the 7th International Conference on Language Resources and Evaluation, LREC- 2010.

Boukhris, F. Boumalk, A. El Moujahid, E., Souifi, H. (2008). La nouvelle grammaire de l'Amazighe. Publications de Institut Royal de la Culture Amazighe.

Chafik, M. (1999). Dictionnaire bilingue : arabe amazighe (tomes 1, 2, 3). Publications de l'Académie Marocaine.

Nejme, F.Z., Boulaknadel, S., Aboutajdine, D.,"Finite State Morphology for Amazighe Language", In Proceedings of the International Conference on Intelligent Text Processing and Computational Linguistics (CICLing), Samos, Greece, 2013.http://dx.doi.org/10.1007/978-3- 642-37247-616

Nejme, F.Z., Boulaknadel, S., Aboutajdine, D. (2013). Analyse Automatique de la Morphologie Nominale Amazighe TALN-RÉCITAL 2013, 17-21 Juin, Les Sables d'Olonne

Nejme, F.Z., Boulaknadel, S., Aboutajdine, D. Formalisation de l'Amazighe standard avec NooJ. Actes de la conférence JEP-TALN-RECITAL, Grenoble, France, 2012.

Nejme, F.Z., Boulaknadel, S., Aboutajdine, D., AmAMorph: Finite State Morphological Analyzer for Amazighe», CIT. Journal of Computing and Information Technology, vol. 24, no. 1, pp. 91-110, 2016.

Oulhaj, L. (2000). Grammaire du Tamazight. Eléments pour une standardization. Centre Tarik ibn Zyad pour les études et la recherche.

Outahajala, M., Zenkouar L., Rosso P. and Martí A. (2010). Tagging Amazighe with Ancora Pipe. In Proceedings of Workshop on LR & HLT for Semitic Languages, 7th International Conference on Language Resources and Evaluation (LREC-2010), Malta, pp. 52-56.

Outahajala, M., Zenkouar L. and Rosso P. (2011a). Building an Annotated Corpus for Amazigh.In Proceedings of the 4th Internatioan Conference on Amazigh and ICT, NTIC-2011, Rabat, Morocco.

Outahajala, M., Benajiba Y., Rosso P. and Zenkouar L. (2011b). POS Tagging in Amazigh using Tokenization and n-gram Character Feature Set. In Proceedings of Symposium Int. sur le Traitement Automatique de la Culture Amazighe (SITACAM-2011), Agadir, Morocco.

Outahajala, M. (2015). Apprentissage supervisé d'un étiqueteur morpho syntaxique automatique de la langue amazighe. Thèsede Doctorat. Ecole Mohammedia d'Ingénieurs, Université Mohamed V-Rabat.

Raiss, H., Cavalli-Sforza, V. Amazigh Nouns Morphological Analyzer. In Proceedings of the 4th Internatioan Conference on Amazigh and ICT, NTIC-2011, Rabat, Morocco.

Kamel, S., Lexique Amazighe de géologie. Rabat, Maroc: IRCAM, 2006.

Sadiqi, F. et Ennaji, M. (2004). A Grammar of Amazigh. Fes: Faculty of Letters, Dhar El Mehraz.

Chaker, S., "Le berbère", Les langues de France (sous la direction de Bernard Cerquiglini), Paris, PUF, pp. 215-227, 2003.

Ouakrim, O., "Fonetica y fonologı´adelBereber", Survey at the University of Autonoma de Barcelona, 1995.

Détection automatique de fin des phrases pour la langue amazighe

Mohamed Biniz[1] **Rachid el Ayachi**[2] **Mohamed Fakir**[3]
[1] Laboratoire de traitement de l'information et aide à la décision (TIAD), Faculté des Sciences and Techniques, Université Sultan Moulay Slimane, Beni Mellal, Maroc.
mohamedbiniz@gmail.com
[2] Laboratoire de traitement de l'information et aide à la décision (TIAD, Faculté des Sciences and Techniques, Université Sultan Moulay Slimane, Beni Mellal, Maroc.
rachid.elayachi@usms.ma
[3] Laboratoire de traitement de l'information et aide à la décision (TIAD Faculté des Sciences and Techniques, Université Sultan Moulay Slimane, Beni Mellal, Maroc.
fakfad@yahoo.fr

Résumé

Cet article présente un système dont l'objectif est de reconnaître les phrases dans un texte, en balisant les phrases et les signes de ponctuation. Il a été testé avec succès sur un corpus de langue amazighe. Le système élaboré se base sur l'algorithme d'Entropie Maximale (MaxEnt) pour extraire les règles. Ces règles sont issues des statistiques, elles-mêmes basées sur l'analyse distributionnelle. L'avantage du traitement effectué ne nécessite aucune connaissance préalable. Sa portabilité et son adaptabilité le rendent précieux pour toute application ou analyse sur le traitement automatique de langage naturel (TALN).

1. Introduction

La plupart des applications de traitements du langage naturel se basent sur le découpage de textes en phrases. Cette tâche est depuis très longtemps automatisée, on parle alors de la reconnaissance de frontières des phrases. La phrase est considérée comme l'unité centrale des processus du traitement du langage naturel. On reconnaît comme phrase la suite des mots qui se trouvent entre des signes de ponctuation tels que le point, le point d'exclamation, le point d'interrogation et d''autres qui précédent ou suivent ces signes. Formellement la ponctuation désigne l'ensemble des signes qui permettent l'interprétation des textes écrits. Traditionnellement la ponctuation est le module de la compétence langagière le plus difficile à maîtriser, d'une part à cause de différents styles appliqués par les auteurs dans la littérature, d'autre part à cause de son caractère écrit. Les marqueurs de ponctuation jouent un rôle important pour indiquer les relations structurelles dans le discours (écrit), par exemple les relations rhétoriques particulières qui résident dans un texte. Des études linguistiques sérieuses sur le rôle de la ponctuation ont commencé dans les années 80 (Chafe, 1988). Zaki et Fawdah (Zaki & Fawdah, 1988) présentent le début d'une théorie linguistique sur la ponctuation où les signes de ponctuation sont des indicateurs de surface pour plusieurs

catégories syntaxiques. Le problème de la détection automatique des phrases se pose à cause de l'ambiguïté de certains signes de ponctuation. Par exemple un point peut être utilisé pour déclarer la fin d'une phrase, mais aussi pour exprimer une abréviation ou un acronyme (ex.: I.R.C.A.M), ou même un nombre décimal, par exemple : 2.14. Les phrases qui suivent montrent l'ambiguïté des signes de ponctuation comme le point :

- ⵔⵉⵅ ⴰⵏ ⵅⵅ ⴰⵎ ⵗⴻⵔⵉⵔ,
- ⴰ ⴰⵛ ⵍⵉⵅ ⴰ ⵣⵣⵎⴰ :

Cette ambiguïté varie selon le type de texte ou de corpus.

Cet article est organisé comme suit: la section 2 donne une brève description d'un ensemble de travaux effectués dans ce domaine, la section 3 présente la méthodologie de l'approche proposée. Ensuite, la procédure d'évaluation intégrée pour tester la performance du système développé est décrite dans la section 4 tandis que la section 5 conclut le travail effectué.

2. Travaux effectués

Dans la littérature, la majorité des travaux ont été concentres sur les langues Arabe(Khalifa, Feki, & Farawila, 2011), Anglaise (Gillick, 2009), Française (Walker, Clements, Darwin, & Amtrup, 2001), Chinoise (Xie, Xu, & Wang, 2012),etc. Les techniques utilisées (souvent des listes grammaticales et de longues listes d'abréviations) visent à reconnaître les cas les plus courants. La majorité de ces dernières sont orientées vers la détection des phrases dans un corpus donné, ce qui rend difficile de les adapter à un autre type de textes ou à une autre langue sans avoir à modifier l'algorithme. De plus, puisque la détection de phrases est juste une première étape dans le traitement automatique du langage naturel, elle ne doit pas demander trop de ressources et ni de calcul au niveau de l'exécution. Donc, le développement d'un système basé sur de grandes listes d'abréviations, ou de lexiques spécialisés avec des informations qui ne s'utiliseront plus par la suite, n'est pas la meilleure solution.

Plusieurs travaux mentionnent l'utilisation d'un détecteur de phrases, mais aucune information donnée sur sa conception ou sur sa performance(Voutilainen, 1995). L'approche la plus simple consiste à l'usage des règles qui reconnaissent des suites de caractères (par exemple : point – espace – lettre majuscule), accompagnées de longues listes d'exceptions pour les abréviations. Une différente approche a été proposée (Rehman & Anwar, 2012): au lieu des listes d'abréviations, une analyse morphologique filtre les mots ayant les mêmes suffixes, pour lesquels la probabilité d'être une abréviation est minime ou nulle. De telles approches demandent plusieurs heures de travail pour construire et renouveler les listes de règles et d'abréviations. De plus, la détection n'est qu'un moyen pour permettre à divers outils d'effectuer leur analyse des phrases découpées, donc le coût de calcul doit être minime. C'est pour cette raison que Reynar et Ratnaparkh (Reynar & Ratnaparkhi, 1997) proposent un modèle basé sur l'entropie maximale qui ne nécessite aucun calcul ni le coût d'une information compliquée. L'information utilisée par ce modèle est fondée sur le segment (token) qui contient le signe de la ponctuation candidate pour la fin d'une phrase, ainsi que les autres segments (tokens), juste avant et juste après celui-là. Pour la construction d'une liste

d'abréviations extraite du corpus annoté, un algorithme simple est utilisé. L'approche de la plus grande entropie dépasse 94% de solution pour le corpus Brown, avec l'utilisation d'une liste manuellement construite d'abréviations, d'appellations et d'acronymes. Si on enlève cette information supplémentaire, le taux est de 92,4%.

Malheureusement, il n'existe aucun travail concernantle texte amazigh écrits en tifinagh jusqu'a la rédaction de cet article. C'est pour cela l'objectif de cet article est de mettre en œuvre un système automatique permettant la détectiondes phrases dans un texte écrit en tifinagh.

3. Méthodologie

Le système élaboré dans le présent travail permet de détecter les phrases d'un texte écrit en tifinagh. Il comporte un ensemble des étapes décrites dans la figure suivante :

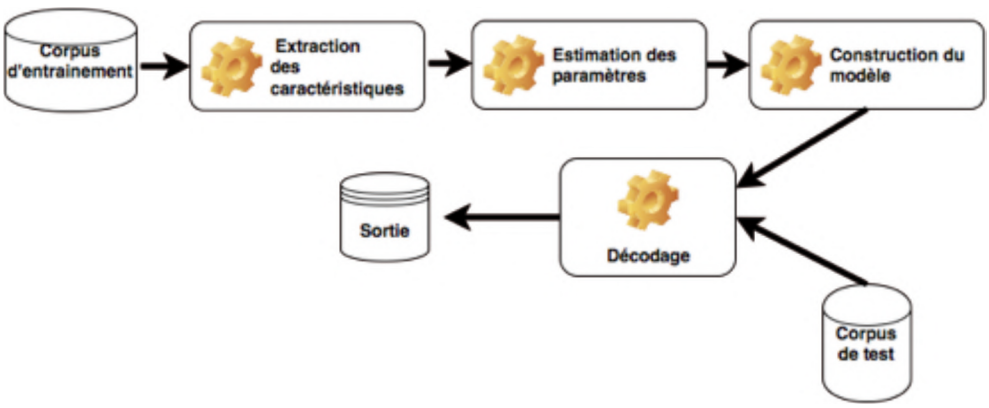


Figure 1 : processus du système

En premier lieu, la phase d'extraction des caractéristiques s'intéresse au dégagement des attributs à partir d'un corpus d'entraînement. Ensuite, les paramètres de Maximum de vraisemblance sont estimés dans la deuxième phase. Après, dans la troisième phase un algorithme d'entropie maximale est adopté pour l'élaboration du modèle. Finalement, la phase de décodage permit de détecter les phrases dans un corpus de test en se basant sur le modèle construit.

Le fonctionnement de ce système est fondé sur le travail de Reynar et Ratnaparkh (Reynar & Ratnaparkhi, 1997) qui se base sur un ensemble de règles:

- L'utilisation des paramètres simples comme la lettre finale d'un mot ou sa longueur dont le coût de calcul est minime.
- L'adoption des règles de désambiguïsation pour les signes de ponctuation.

Ces règles sont extraites à partir des informations tirées des statistiques sur des corpus non annotés et sont applicables pour le texte amazigh écrits en tifinagh.

Avant de procéder à la détection des frontières des phrases d'un texte, nous définissons un ensemble de segments (tokens), par exemple : caractères, nombres, signes de ponctuation. Ces signes se diffèrent d'une langue à une autre, à part quelques signes comme le point final. Le point d'exclamation, qui est commun pour les deux langues Anglais et Tifinagh. La procédure de la segmentation suit la règle suivante : deux segments sont séparés par des espaces. Un segment qui est suivi par un des signes de ponctuation décrits ci-dessous est considéré comme fin probable de la phrase (en respectant la forme particulière de chaque langue).

- Point.
- Point d'exclamation !
- Point d'interrogation ?
- Points de suspension …

Les corpus sont composés de textes des styles divers :poèmes, magazines, littérature, etc., et ils sont non annotés. Le système n'a pas besoin d'une connaissance préalable ou d'un dictionnaire pour fonctionner. Il faut préciser que le corpus Tifinaghutiliséatteint les 4010mots.

3.1. Caractéristique dele texte amazigh écrits en tifinagh

En effet, il n'y a pas de différences quant à l'utilisation de la ponctuation dans le texte amazigh écrits en tifinaghau sein de notre recherche. Les différences se résument à la graphie et aux particularités morphologiques que nous décrivons brièvement. Les particularités morphologiques du texte amazigh écrits en tifinagh ont été prises en compte (En traitant les textes codés en Unicode) afin de finaliser l'algorithme de la détection de phrases et désambiguïser l'utilisation de signes de ponctuation. Le mot tifinagh graphique est facile à identifier : c'est ce qui s'écrit en un seul bloc entre deux blancs. En Tifinagh, il n'y a pas de distinctionentreles minuscules etles majuscules. Ceci est intéressant pour la reconnaissance des acronymes, car nous pouvons les identifier suivant les mêmes règles que pour les langues occidentales : «Alphabet - Point - Alphabet - Point», exemple etc.

Il faut juste mentionner que pour le tifinagh le signe de ponctuation point-virgule (;) correspond au point et signifie donc la fin d'une phrase (exemple : ⵄⵍⵍⴰ ⵠⵣⵓⵠ ⵅ ⵡⵍⵉⵓⴰ;). Il faut aussi mentionner que le double point « : », étant donné qu'il peut signifier la fin d'une proposition, est un marqueur de fin d'une phrase. Ceci est le caspour plusieurs langues comme l'Arabe, l'Anglais etle Français.

3.2. Prédicats contextuels

À ce niveau, Nous allons décrire le choix des paramètres pour la reconnaissance des fins des phrases, sachant que les informations équivalentes ont été extraites des autres langues testées tels l'Anglais et l'Arabe. La partie du candidat qui précède la frontière de la phrase potentielle est appelée préfixe et la partie suivante représente le suffixe. Notre système vise à maximiser la performance en utilisant le « modèle » suivant :

- Le préfixe
- Le suffixe
- Que ce soit le préfixe ou le suffixe est sur la liste des abréviations induites
- La gauche du candidat
- Le droit du candidat
- Que ce soit le mot à gauche ou à droite du candidat figure sur la liste des abréviations induites.

Les informations de ce modèle sont appliquées dans l'exemple suivant :

ⵍⵢⵔⵙⵜⵜⴰⵏⵜ, ⵔⵜⵉⵎⵣⵡⵓⵔⵜ.

Exemple 1: exemple d'une phrase

Cela génère ces résultats : PreviousWord = ⵡ, Following Word = ⵔⵜ, Prefix = ⵔⵙⵜⵔⵜ, Suffixe = NULL, Prefix Feature = NULL.

Le système utilise seulement l'identité du candidat et de ses mots voisins, et une liste des abréviations induites à partir des données d'entraînement. La liste des abréviations est produite automatiquement à partir des données d'entraînement, et les contextuelles sont également générées automatiquement par la numérisation des données d'entraînement. Par conséquent, pas de règles où des listes faites à la main sont requises par ce système et il peut être facilement reformé pour d'autres langues ou genres de texte.

3.3 Modèle d'entropie maximale

Les résultats du modèle de probabilité(Reynar & Ratnaparkhi, 1997) sont « oui » et « non », où « oui » dénote qu'une frontière de phrase potentielle, et « non »indique qu'il n'y a pas une limite d'une phrase réelle.

Pour chaque jeton de frontière de phrase potentiel (.;.! et ?), nous tenons à estimer une distribution conjointe de probabilité p de celui-ci et son contexte environnant qui se considère comme une phrase réelle de la distribution de probabilité frontière.On procède par l'utilisationd'un modèle d'entropie maximale définitpar l'équation (1) :

$$p(a|b) = \frac{1}{Z(b)} \prod_{j=1}^{k} \alpha_j^{f_j(a,b)} \qquad (1)$$

Sachant que k est le nombre de caractéristiques et Z(b) est un facteur de normalisation pour faire en sorte que $\sum_a p(a|b) = 1$ .Z(b) est défini par la formule suivante :

$$Z(b) = \sum_{a} \prod_{j=1}^{k} \alpha_j^{f_j(a,b)} \qquad (2)$$

Les prédicats contextuels jugés utiles pour la détection de frontière de phrase, que nous avons décrits plus haut dans l'exemple 1, sont codés dans le modèle en utilisant des fonctionnalités. Par exemple, une fonction utile pourrait être :

$$f_j(a,b) = \begin{cases} 1 & si\ Prefix\ (b) = |\ et\ a = no \\ 0 & sinon \end{cases} \qquad (3)$$

Cette fonctionnalité permettra au modèle de découvrir que le mot ⵅ se produit rarement comme une frontière de phrase. Par conséquent, le paramètre correspondant à cette fonction, nous l'espérons stimuler la probabilité si le préfixe est ⵅ. toutes les fonctions qui se produisent 10 fois ou plus dans les données d'entraînement sont conservées dans le modèle, et les paramètres du modèle sont estimés avec l'algorithme « Generalized itératif Scaling » (Darroch & Ratcliff, 1972).

3.4 Algorithme Generalized iteratif Scaling

L'algorithme « Generalized itératif Scaling » (Darroch & Ratcliff, 1972), ou GIS nous permet d'estimer les paramètres de p* (maximum likelihood). Il exige que la somme des caractéristiques estégale à une constante C pour tout (a, b) ∈ A x B, comme nous montre l'équation (4) :

$$\sum_{j=1}^{k} f_j(a, b) = C \qquad (4)$$

Dans le cas où cette condition n'est pas vérifiée, nous utilisons l'ensemble de l'entraînement pour choisirC, qui est définit par l'équation (5) :

$$C = \max_{a \in A,\, B \in T} \sum_{j=1}^{k} f_j(a, b) \qquad (5)$$

Ensuite,on ajoute une « correction »exprimée par l'équation (6):

$$f_l(a, b) = C - \sum_{j=1}^{k} f_j(a, b) \quad avec\ \ l = k + 1 \qquad (6)$$

Pour toute (a,b) paire, on note que fl(a, b) est comprise entre 0 et C, où C peut être supérieur à 1. En théorie, une constante de correction pour exiger la contrainte de l'équation (5) pour

toutes les (a, b) paires devrait être dérivée de l'espace des possibles événements A x B. Cependant, un des ensembles de l'entraînement est généralement précis en pratique.

Algorithm 1 (GIS):

• Initialisation :

$$\alpha_j^{(0)} = 1$$

• Iteration jusqu'à la convergence :

$$\alpha_j^{(n+1)} = \alpha_j^{(n)} \left[\frac{E_{\tilde{p}} f_j}{E_{p(n)} f_j} \right]^{\frac{1}{C}}$$

Tel que :

$$E_{p(n)} f_j = \sum_{a,\, b} \tilde{p}\,(b)\, p^{(n)}\,(a|b)\, f_j\,(a,\, b) \qquad (7)$$

$$p^{(n)}\,(a|b) = \frac{1}{Z(b)} \prod_{j=1}^{l} (\alpha_j^{(n)})^{f_j\,(a,\, b)}$$

4. Résultats expérimentaux :

Pour évaluer notre système, nous avons adopté trois mesures à savoir : Précision, Rappel et F-mesure (Powers, 2011).

La précision est définie comme le nombre total de limites de phrases correctement extraites sur le nombre total de limites extraites par le système.

$$Precision = \frac{Nombre\ total\ de\ limites\ de\ phrases\ correctement\ extraites}{Nombre\ total\ de\ limites\ extraites} \qquad (8)$$

Rappel est défini dans l'équation (9) comme étant le nombre total de limites de phrases correctement extraites sur les véritables limites du nombre total existant dans la collection.

$$Rappel = \frac{Nombre\ total\ de\ limites\ de\ phrases\ correctement\ extraites}{Nombre\ total\ de\ limites\ de\ phrases\ correctes\ existant} \qquad (9)$$

F-mesure représente la moyenne harmonique de précision et de rappel, ce qui est un bon indicateur d'un certain nombre de performances des systèmes.

$$F - measure = \frac{2 * Précision * Rappel}{Précision + Rappel} \qquad (10)$$

Pour la partie expérimentale, nous utilisons l'implémentations d'OpenNLP (Reese, 2015) version 1.6.3 avec une adaptation spécifique pour supporter le texte amazigh écrits en tifinagh.

Le Tableau suivant montre les résultats obtenus pour les trois mesures adoptées :

Corpus d'entrainement	Corpus de test	Précision	Rappel	F-mesure
20	40	1.00	1.00	1.00
30	30	1.00	1.00	1.00
40	20	1.00	1.00	1.00

Tableau 1: Résultats obtenus par le système

Le corpus utilisé contient en total 60 phrases de différentes ponctuations. Ce corpus est divisé en deux parties :

- Corpus de test
- Corpus d'entraînement

Les résultats figurés dans le Tableau 1 montrent clairement la performance de notre système élaboré.

5. Conclusion

La procédure de détection des phrases dans un corpus donné constitue une étape primordiale pour toute application TALN.

Dans cet article, nous avons développé un système robuste qui permet de reconnaître les frontières d'une phrase dans un texte en tifinagh. Ce système comporte quatre étapes : l'extraction des caractéristiques, l'estimation des paramètres, l'élaboration du modèle et le décodage.

L'avantage de notre système c'est qu'il n'a pas besoin d'aucune connaissance préalable et aucun corpus annoté.

Pour la validation du système, un ensemble de tests expérimentaux ont étéeffectués avec succès sur un corpus de tifinagh construit localement. Les résultats trouves jugent la performance du système.

Références

Chafe, W. (1988). Punctuation and the Prosody of WrittenLanguage. Written Communication, 5(4), 395–426.

Darroch, J., &Ratcliff, D. (1972). Generalized Iterative Scaling for Log-Linear Models. The Annals of MathematicalStatics, 43(5), 1470–1480.

Gillick, D. (2009). Sentence boundarydetection and the problemwith the US. In Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for ComputationalLinguistics, Companion Volume: Short Papers (pp. 241–244). Association for ComputationalLinguistics.

Khalifa, I., Feki, Z., &Farawila, A. (2011). Arabicdiscourse segmentation based on rhetoricalmethods. Int. J. Electric Comput. Sci, 11(1).

Powers, D. M. (2011). Evaluation: fromprecision, recall and F-measure to ROC, informedness, markedness and correlation.

Reese, R. M. (2015). Natural Language Processingwith Java. Packt Publishing Ltd.

Rehman, Z., & Anwar, W. (2012). A hybridapproach for urdu sentence boundarydisambiguation. Int. Arab J. Inf. Technol., 9(3), 250–255.

Reynar, J. C., & Ratnaparkhi, A. (1997). A maximum entropyapproach to identifying sentence boundaries. In Proceedings of the fifthconference on Appliednaturallanguageprocessing (pp. 16–19). Association for ComputationalLinguistics.

Voutilainen, A. (1995). NPtool, a detector of English noun phrases. Proceeding of Workshop on Very Large Corpora:Academic and Industrial Perspectives,.

Walker, D. J., Clements, D. E., Darwin, M., & Amtrup, J. W. (2001). Sentence boundarydetection: A comparison of paradigms for improving MT quality. In Proceedings of the MT Summit VIII.

Xie, L., Xu, C., & Wang, X. (2012). Prosody-based sentence boundary detection in chinese broadcast news. In Chinese Spoken Language Processing (ISCSLP), 2012 8th International Symposium on (pp. 261–265). IEEE.

Zaki, A., & Fawdah, ʻAbd al-Raḥman ibn Ibrahim Ibn. (1988). الترقيم وعلاماته في اللغة العربية. Maktabat al-Tawʻīyah al-Islāmīyah li-Iḥyā’ al-Turāth al-Islāmī.

TICAM'16, p. 137 - 148

Le Machine Learning : numérique non supervisé et symbolique peu supervisé, une chance pour l'analyse sémantique automatique des langues peu dotées

Hammou Fadili[1,2]
[1] Laboratoire CEDRIC du Conservatoire National des Arts et Métiers de Paris
192, rue Saint Martin, 75141, Paris cedex 3, France
[2] (Pôle Systèmes d'Information et du Numérique, Programme Maghreb) de la FMSH Paris
190, avenue de France 75013, Paris, France
Hammou.fadili@cnam.fr / fadili@msh-paris.fr

Résumé

Les données non structurées dominent l'univers de la production et de la publication des données, et en représentent, d'après plusieurs études, plus de 80%. Ce type de contenus, constitue la partie riche et précieuse en termes de données, d'informations et de connaissances; donc nécessaire à intégrer et à prendre en considération dans les processus d'analyses et d'exploitations des données. L'analyse des données non structurées reste une discipline difficile, car elle repose sur de nombreux (pré-) traitements numériques (formalisation, normalisation, corpus, annotations, etc.) de la langue naturelle, faisant souvent défaut, surtout dans le cas des langues peu dotées. Dans cet article nous présentons une approche, bien adaptée aux cas où on ne dispose pas ou que de peu de données traitées. Elle est basée sur des méthodes d'apprentissagesnumériques non supervisés indépendants de la langue et symboliques peu supervisés, permettant d'exploiter directement des données brutes ou seulement des petites quantités de données traitées, comme base d'apprentissage pour l'interprétation des données. En l'appliquant à un cas concret d'une langue peu dotée, nous avons pu, montrer l'utilité et surtout l'opportunité que ces technologies pourraient constituer pour contourner les problèmes dont souffre ce type de langues, facilitant ainsi leur accès dans le monde de l'analyse sémantique automatique des données non structurées. Cette étude a été validée à travers des expérimentations confirmant de bons résultats pour l'approche.

1. Introduction

D'une manière générale, l'analyse et l'interprétation sémantiques des données textuelles, n'est pas encore totalement prise en charge par les systèmes actuels. Ceci est dû aux problèmes et aux difficultés liées à la modélisation et à la formalisation de la complexité de la langue naturelle dans sa globalité, à ses très nombreux cas possibles, à ses nombreuses exceptions,

etc. Cesdifficultés sontencore plus accentuées, dans le cas des langues peu dotées, faute de ressources informatiques et de corpus traités suffisants.

Or, dans le domaine de l'analyse sémantique des données, il existe plusieurs technologies indépendantes de la langue traitée,ou ayant des capacités de généralisation à partir de petites bases d'apprentissage. Elles sont généralement basées, sur des notions d'apprentissages numériques non supervisés et/ousymboliques peu supervisés, ne nécessitent pas / ou peu de (pré)traitements préalables.Ces méthodes sont donc applicables à n'importe quelle langue, constituant ainsi une chance pour l'analyse sémantique automatique des langues peu dotées.

Afin de valider ces approches, nous avons testé des technologies, déjà étudiées pour le Français et l'Anglais, sur une langue peu dotée. Ceci afin de montrer, d'une part, leur indépendance par rapport aux langues et d'autre part, la possibilité de leur application à des langues peu dotées. Pour cela, nous avons exploité des dictionnaires existants pour enrichir et instancier une première partie du modèle de données proposé pour l'apprentissage. L'autre partie a été complétée en exploitant des technologies basées sur des apprentissages non supervisés : LSA (Latente SemanticAnalysis) pour la génération de la sémantique latente et LDA (Latent Dirichlet Allocation) pour la génération des thèmes du contexte général. D'autres outils classiques pour gérer le POS, effectuer le streaming, générer le contexte local, etc., ont été également exploités.

Dans cet article, la première partie sera consacrée à rappeler les motivations de ce travail, la deuxième partie sera consacrée à la description de notre contribution : la définition du modèle d'apprentissage, la préparation des données, présentation des technologies exploitées et/ou améliorées, description de l'architecture du système, ainsi que son évaluation. La dernière partie conclura l'article.

2. Motivation

Dans ce paragraphe, nous allons rappeler quelques éléments sur l'analyse sémantique des données non structurées d'une manière générale. Ceci, afin de montrer les difficultés qu'on rencontre dans le cas des langues peu dotées et les solutions qui peuvent être adoptées.

2.1. Analyse des données textuelles

Le processus d'analyse des données est généralement composé de deux phases :

- Phase de préparation des données
- Phase de l'analyse sémantique

Phase de préparation des données (filtrage des données)

Cette phase correspond aux prétraitements linguistiques des données textuelles, nécessaires, aux traitements des couches supérieures dont celles du traitement automatique de la sémantique. A ce niveau, le processus consiste en général, en l'analyse et en la génération d'un réseau de mots et de relations, qu'on appelle un réseau morphosyntaxique, dépourvu de

sens, représentant seulement les éléments constituant du texte. La génération d'un tel réseau de mot suit en « général » le processus suivant :

♦ Segmentation

- Tokénisation (tokenization) : découpage du texte en unités lexicales élémentaires (tokens)
- Segmentation (Sentence Segmentation) : découpage du texte en phrases

♦ Analyse morphologique

- Racinisation (stemming): recherche de la racine
- Lemmatisation (lemmatization) : regroupement des formes de mots qui appartiennent à la même famille

♦ Analyse syntaxique

- Etiquetage grammatical (Part-of-speech tagging : POS) : détermine la catégorie lexicale (nom, verbe, adj, adv, etc.), annotation des catégories grammaticales
- Analyse de groupes grammaticaux (Chunking) : détermine les groupes grammaticaux (nominaux, verbaux, etc).

Phase d'analyse sémantique

Cette phase exploite les résultats de la phase précédente et des traitements sophistiqués pour la déduction de la sémantique dans le contexte. Ceci complète le processus général de l'analyse textuelle :

♦ Analyse sémantique

- Exploitation des étapes précédentes du processus et des modèles symboliques ontologiques et/ou mathématiques statistiques pour l'analyse du sens.

Cette phase sera décrite d'une manière détaillée, à travers notre approche, un peu plus tard dans cet article. Les résultats d'analyses des deux phases sont en général stockés sous formats structurés comme en xml, csv, etc., échangeables et exploitables par des applications tierces.

2.2. Cas des langues peu dotées

D'une manière générale, les langues peu dotées, sont des langues qui souffrent de plusieurs problèmes liés à leur traitement automatique : problèmes liés à la graphie, au manque d'un système d'écriture stable, au manque de ressources informatiques et linguistiques. Le manque de ressources langagières concerne les dictionnaires, thésaurus, corpus traités, etc.; le manque d'outils numériques concerne les outils du traitement automatique de la langue naturelle : analyseurs morphologiques, syntaxiques, sémantique, etc.Tous ces éléments rendent difficile,voire impossible l'analyse sémantique des langues peu dotées.

Afin de contourner ces problèmes, nous proposons l'exploitation, de solutions,indépendantes / peu dépendantes de la langue traitée et ayant fait leur preuve pour d'autres langues mieux dotées (Anglais, Français, etc.).

3. Notre approche

Notre approche permet de contourner les problèmes rencontrés pour les langues peu dotées. Elle est basée sur des apprentissages des modèles de représentation pour le traitement automatique de la sémantique du texte. D'une manière générale, il y a les approches dites numériques qui exploitent les fréquences des mots et les modèles mathématiques statistiques et probabilistes ; puis il y a les approches dites symboliques qui exploitent la structure des données, leur description et des systèmes de règles. Dans le premier cas, un texte est représenté dans un espace vectorielle, dont les dimensions sont des termes sélectionnés à partir d'un vocabulaire donné (un sous-ensemble des mots du texte), par un vecteur V des fréquences des mots du texte. Dans le deuxième cas, le texte est représenté par les mots le constituant ainsi que des propriétés et des règles issues des différents traitements linguistiques ou des formalisations spécifiques telles les ontologies. D'une manière générale, les modèles symboliques exploitant les données d'experts sont efficaces, mais sont difficiles à mettre en place, sauf pour des domaines très restreints. La difficulté réside dans le fait qu'on ne peut pas modéliser et décrire d'une manière complète la totalité d'un domaine donné. Cela demande beaucoup de ressources et de connaissances, donc beaucoup de travail. Les méthodes statistiques quant à elles n'ont pas besoin de toutes ces informations pour spécifier et décrire un modèle d'analyse ; elles exploitent seulement les données d'analyse et des modèles mathématiques. Dans notre cas, nous avons combiné les deux méthodes : on exploite des techniques d'apprentissage non supervisé pour préparer une partie ou la totalité des données qu'on combine avec un petit échantillon de données sur lequel on applique des techniques d'apprentissage peu-supervisé.

3.1. Contexte de l'étude

Cette étude s'inscrit dans le contexte général, de l'analyse sémantique des données textuelle des langues peu dotées.Elle est basée sur des apprentissages non/peu supervisés, et a été exploitée pour la détection dela sémantique dans la langue Amazigh. Pour se faire, on a juste traduit un mini corpus du Français sur lequel on a appliqué les différents algorithmes. Dans ce qui suit, une description des éléments du processus, exploités et/ou améliorés pour les besoins de l'étude.

3.2. Dictionnaires

En terme de dictionnaires, le manque de corpus annotées, d'une part, et le soucis de rendre notre démarche générique (sans se limiter à un domaine particulier), d'autre part, nous ont obligé à mettre en place une stratégie basée sur l'exploitation d'un dictionnaire de langue et des apprentissagesautomatiques. On a exploité des méthodes d'apprentissages mixtes (peu supervisés et non supervisés) et des technologies de prétraitements (préparation) des données, pour instancier le modèle étendu de données et construire les bases de validation, d'entrainement et de test.

3.3. Contexte sémantique global

La notion de contexte sémantique dont on parle ici, concerne tous les paramètres pouvant influencer le sens des mots dans le texte. Ce contexte peut être considéré comme une agrégation de paramètres appartenant aux ensembles suivants :

- contexte textuel : le contexte du texte, domaine, thème, univers du discours,…
- contexte local : l'ensemble des mots situés de part et autre du mot, on l'appelle aussi la fenêtre textuelle
- contexte linguistique : le rôle du mot dans la phrase
- contexte d'utilisation : le domaine d'utilisation, but d'utilisation…
- contexte de l'auteur : les sens particuliers des mots suivant certains auteurs
- contexte du lecteur (point de vue d'analyse) : la compréhension particulière du sens de certains mots par certains lecteurs
- tous autres paramètres pouvant influencer le sens des mots.

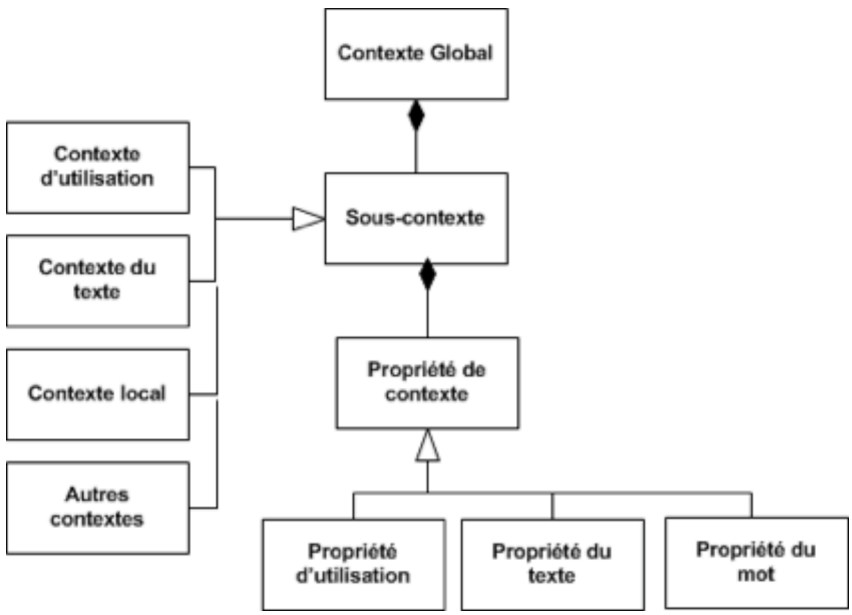


Figure. 1 : Méta-modèle du contexte global

Ces éléments définissant le contexte sémantique global, sont au moins de trois types :

- Ceux contrôlés par l'utilisateur : domaine, but…
- Ceux véhiculés par le texte : thèmes, contexte textuel, local, linguistique…

- Ceux pour lesquels on peut faire des choix (cibler une lecture particulière du texte), i.e. qu'on peut fixer au départ ou extraire du texte lui-même : lecteurs, auteurs…

Des ontologies peuvent être utilisées pour modéliser et représenter des éléments des contextes. La partie du contexte connue à l'avance, doit être instanciée par l'utilisateur dès le départ ; l'autre partie dépendante du texte, doit être calculée et générée automatiquement au moment de l'analyse (contexte textuel, local, etc.).

3.4. Relations sémantiques

On considère ici toutes les relations influençant le sens que peuvent entretenir deux mots. On peut citer :

- les relations paradigmatiques : les relations de sens que peuvent entretenir deux mots
- les relations syntagmatiques : les relations de collocation modifiant le sens des mots
- les relations thématiques : les relations que peuvent entretenir des mots par rapport à une thématique
- les relations de variation : relations entre un mot et ses variations
- les relations racinisation : relation entre un mot et sa racine
- les relations sigles : relations d'équivalences entre les entités et les sigles, les initiales, et autres annotations…
- toutes les autres relations utiles pour la détection du sens des mots

En s'inspirant d'un travail de Igor Mel'cuk(Mel'cuk, 1997), nous avons recensé une soixantaine de relations sémantiques (Fadili, 2013).

Syn, Anti, Conv$_{ij}$*, Contr, Epit, Gener, Figur, S*$_0$*, A*$_0$*, V*$_0$*, Adv*$_0$*, S*$_i$*, S*$_{instr}$*, S*$_{med}$*, S*$_{mod}$*, S*$_{loc}$*, S*$_{res}$*, Able*$_i$*, Qual*$_i$*, Sing, Mult, Cap, Equip, A*$_i$*, Adv*$_i$*, Imper, Result, Centr, Magn, Plus, Minus, Ver, Bon, Pos*$_i$*, Loc*$_{in}$*, Loc*$_{ab}$*, Loc*$_{ad}$*, Instr, Propt, Copul, Pred, Oper*$_i$*, Func*$_i$*, Labor*$_{ij}$*, Incep, Cont, Fin, Caus, Perm, Liqu, Real*$_i$*, Fact*$_i$*, Labreal*$_{ij}$*, Involv, Manif, Prox, Prepar*$_i$*, Degrad, Son, Obstr*$_i$*, Stop*$_i$*, Excess*$_i$*, Sympt*$_{ijk}$.

Figure 2 : Relations sémantiques

Dans le cas des relations sémantiques, on a besoin de dictionnaires modélisant les différents liens de sens entre les mots. On peut se baser sur des dictionnaires existants pour représenter une partie des relations. Ces dictionnaires souvent incomplets, doivent êtres enrichis automatiquement en utilisant des technologies basées sur des statistiques et des règles d'apprentissage ; afin de détecter de nouvelles relations, nécessaires à l'amélioration de l'interprétation automatique.

3.5. Modèle de données pour l'apprentissage

Pour l'analyse du texte, nous avons besoin d'une représentation la plus complète possible d'un mot et d'un texte, qui permet de capter toutes les informations discriminantes et de déduire le sens exact du mot dans le texte. Dans le présent travail, nous proposons un modèle

exprimant essentiellement deux notions essentielles, la notion de contexte et la notion des relations sémantiques. Nous supposons, que la définition la plus large possible du contexte et la détermination de tous les mots liés d'un mot permettent de déterminer d'une manière non ambiguë son sens réel.

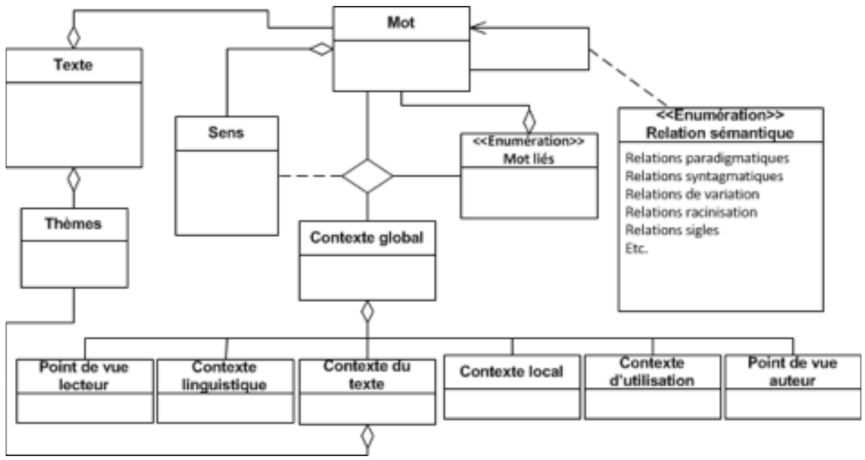


Figure 3 : Modèle de données pour l'apprentissage

3.6. Sémantique latente pour l'instanciation d'une partie du modèle d'apprentissage

Dans ce paragraphe, nous allons présenter des approches statistiques pour l'extraction de la sémantique latente. Ces approches ont été exploitées pour compléter l'instanciation du modèle des données étendu et du contexte globale.

LSA : Analyse sémantique latente (Latente SemanticAnalysis, LSA) (Deerwester *et al.* 1990) consiste à découvrir de manière statistique la sémantique latente dans un corpus. Elle permet d'établir des relations entre les termes (termes en plus de leur contexte) en exploitant une matrice de cooccurrences. Grâce à LSA, on peut exprimer par exemple une relation de synonymie entre deux mots qui apparaissent souvent ensemble et une relation de polysémiesi un mot apparait souvent dans plusieurs contextes différents. L'obtention de la sémantique latente s'obtient par la construction de la matrice approchée X_k de la matrice des cooccurrences X (apparition des mots du texte dans les différents contextes locaux) qu'on décompose en une matrice diagonale de valeurs singulières et en deux matrices orthogonales $X = U\Sigma V^T$, puis qu'on réduit au nombre des valeurs singulièresk non nulles $X_k = U_k \Sigma_k U_k^T$.

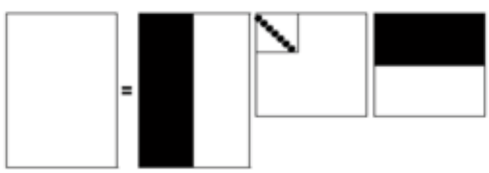


Figure 4 : Réduction des dimensions dans LSA

X_k ne contient que la sémantique des mots la plus importante d'un point de vue statistique pour le document (la sémantique latente des mots dans le corpus). Cette matrice peut être utilisée pour calculer la distance entre les termes ou entre les contextes locaux, en utilisant la distance du cosinus entre les vecteurs lignes et colonnes respectivement. L'intersection d'une ligne et d'une colonne détermine la pertinence du terme associé à la ligne par rapport au contexte associé à la colonne.

LDA : Allocation latente de Dirichlet (Latent Dirichlet Allocation, LDA) (Blei & Lafferty, 2009) consiste à découvrir les thèmes latents d'un corpus correspondant à une distribution spécifique de mots fréquemment groupés. LDA fait partie d'une catégorie de modèles appelés "topic models", qui cherchent à découvrir des structures thématiques cachées dans les textes se présentant comme un mélange de thèmes. La répartition des mots par rapport aux thèmes se fait de la manière suivante : après avoir instancié le nombre de thèmes et une répartition aléatoire de tous les mots sur tous les thèmes, on ajuste l'appartenance des mots aux thèmes, en calculant les probabilités p(theme|document) * p(mot|theme) pour tous les thèmes et pout tous les mots. Dans notre cas, LDA a été utilisée pour associer un contexte à un document à partir des mots qui le composent. Elle a été également utilisée pour établir des liens entre les mots et les thèmes. Ces deux notions ont été exploitées pour augmenter la caractérisation du modèle des données pour l'apprentissage (décrit un peu plus haut dans ce document).

Ces deux méthodes sont non-supervisées et ont souvent besoin d'être employées avec une décomposition en valeur singulière, pour réduire leur dimensionnalité, éviter les matrices creuses, et conserver un maximum de la significativité.

3.7. Apprentissages du sens

L'approche d'apprentissagesdéfinie et exploitée pour l'interprétation et la déduction du sens des mots dans le texte, est basée sur un réseau profond, permettant de modéliser avec un haut niveau d'abstraction des modèles de données articulés autour de transformations non linéaires.On l'a conçu pour permettre des apprentissages mixtes : les premières couches ou couches basses utilisent des apprentissages non supervisés et les dernières couches des apprentissages supervisées ou peu supervisées (couches de décision). En d'autres termes, les premières couches ont été consacrées à l'extraction des caractéristiques à partir des informations latentes (apprentissage non supervisé) et les couches suivantes à la prise de décision (apprentissage peu supervisé).

3.8. Architecture

Le processus consiste en deux étapes principales :

- préparation des données,
- application des algorithmes d'apprentissage.

La première étape est elle-même composée en deux phases :

- La première phase consiste à générer les entrées suivant le modèle des données. Le but est de créer un vecteur par mot, enrichi par des propriétés discriminantes du sens. Pour cela, on exploite les technologies du POS, tokenisation, stemming et spliting pour

générer la catégorie grammaticale et le contexte local, puis les technologies de gestion de la sémantique latente LSA etLDApour générer le contexte du texte (thèmes) et certaines relations sémantiques latentes entre les mots.

- La deuxième phase consiste à exploiterun dictionnaire de langue pour extraire tous les Synset du mot, et enrichir le vecteur généré dans la première phase par, la description du Synset (gloss=définition+exemple) et tous les mots liés par les relations sémantiques. Au bout de cette deuxième phase, on obtient un vecteur globalement contextualisé et enrichi sur lequel on peut appliquer les algorithmes de classification et déduire le sens réelle, comme sortie du système.

Figure 5 : Architecture

3.9. Evaluation

Nous avons exploité les fonctionnalités des outils Automap et Weka pour générer un fichier .arff (format exigé par Weka) d'instances suivant le modèle de données et pour implémenter un perceptron multicouche pour la détection du sens. Afin de rendre le système « honnête : les données de validation différentes de celles de test», nous avons séparé les données générées en trois parties comme suivant :

- 30% des données pour la validation, ceci afin d'optimiser les hyper-paramètres du système : le pas d'apprentissage, le type de la fonction d'activation et le nombre de couches.
- 60% pour l'entrainement, ceci afin d'estimer les meilleurs coefficients (w_i) de la fonction du réseau de neurones, minimisant l'erreur entre les sorties réelles et les sorties désirées.
- et 30% pour les tests, ceci afin d'évaluer les performances du système.

La génération du fichier .arff, s'est fait grâce, dans un premier, à la plateforme Automap, qui nous a permis d'instancier une partie du modèle (POS, Contexte local, etc.) via un fichier .csv. Puis dans un deuxième temps, grâce à la plateforme Weka, qui nous a permis d'appliquer sur le texte les technologies non supervisées (LSA et LDA) pour générer l'autre partie du modèle des données. La plateforme Weka a ensuite été utilisée pour l'implémentation d'un réseau de neurones de type « perceptron multicouche » appliqué aux instances pour l'évaluation.

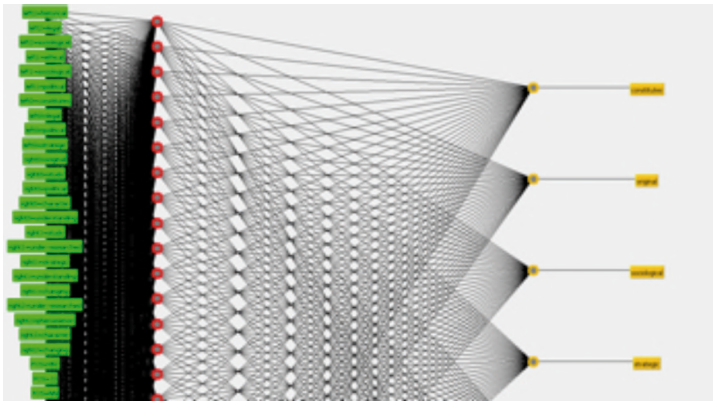


Figure 6 : Application du perceptron multicouche

Ci-après quelques résultats des tests :

Correctly Classified Instances	71.4286%
Incorrectly Classified Instances	28.5714%

Tableau 1: Vue d'ensemble

	Precision	Recall	F-Measure	ROC Area	Class
	0	0	0	0.917	iyya
	1	1	1	1	anezli
	1	1	1	1	tasnametti
	1	1	1	1	astrategic
	0.5	1	0.667	0.917	tisfrass
	0.5	1	0.667	0.917	yann
	0	0	0	0.917	azamul
WeightedAvg.	0.571	0.714	0.619	0.952	

Tableau 2 : Mesures

a	b	c	d	e	f	g	<-- classified as
0	0	0	0	0	1	0	| a=iyya
0	1	0	0	0	0	0	| b=anezli
0	0	1	0	0	0	0	| c=tasnametti
0	0	0	1	0	0	0	| d=astrategic
0	0	0	0	1	0	0	| e=tisfras
0	0	0	0	0	1	0	| f=yann
0	0	0	0	1	0	0	| g=azamul

Tableau 3 : Matrice de confusion

Les résultats des tests montrent les bonnes performances de l'approche. Le système à réussit grâce à l'apprentissage sur une partie des instances à déduire le sens réel de tous les mots qui étaient mal classés au départ.

4. Conclusion

Cet article propose une approche d'apprentissage indépendante de la langue traitée, pour instancier le modèle de données d'apprentissage et lui appliquer des méthodes d'apprentissages peu supervisés. Elle a l'avantage de contourner les problèmes majeurs rencontrés dans l'analyse des données non structurées dans le contexte des langues peu dotées, à savoir le manque d'outils et de données annotées, sémantiquement et automatiquement exploitables. Les expériences menées sur des exemples confirment d'une part l'apport considérable du modèle proposé pour la détection du sens réel des mots dans le texte et d'autre part,l'application de l'approche sur les langues peu dotées.

Références

Akoka J. *et al.*, 2014. A Semantic Approach for Semi-Automatic Detection of Sensitive Data. Information Resources Management Journal, vol. 27, n° 4, pp. 23-44.

Blei D. et Lafferty J., 2009. Topic Models.In A. Srivastava and M. Sahami, editors, Text Mining, Classification, Clustering, and Applications . Chapman & Hall/CRC Data Mining and Knowledge Discovery Series.

Cybenko, G., 1989. Approximation by Superpositions of a Sigmoidal Function. Math. Control Signals Systems, 2, 303-314.

Deerwester, S. *et al.*, 1990.Indexing by latent semantic analysis.Journal of the American Society for Information Science, vol. 41, no. 6, pp. 391-407.

Fadili H., 2013. Towards a new approach of an automatic and contextual detection of meaning in text, Based on lexico-semantic relations and the concept of the context., IEEE-AICCSA, Ifrane (Morroco), May.

Hornik, K., 1991. Approximation capabilities of multilayer feed for ward nets. NeuralNetworks. 4, 231–242.

Rumelhart, D., Hinton G., Williams R., 1986.Learning internal representationby error propagation.In Parallel Distributed Processing, Explorations in the Microstructure of Cognition. Cambridge, MA: MIT Press ,Vol. 1, pp. 318–362.

Turney D., 2006. Similarity of semantic relations. Computational Linguistics, 32(3):379–416.

TICAM'16, p. 149 - 157

Etat de l'art de la traduction automatique des langues – Approches & Méthodes

Naoual Bouhrim[1] **Lahbib Zenkouar[2]**
[1] Département LEC EMI – UNIVERSITE MOHAMMED V
naoualensias@gmail.com
[2] Département LEC EMI – UNIVERSITE MOHAMMED V
Lahbib.zenkouar@gmail.com

Résumé

Cet article présente l'état d'art des travaux et recherches effectués dans le domaine de la traduction automatique des langues (TAL) en citant les principales approches à savoir : Les approches classiques, approches computationnelles et les approches probabilistes.

Un corpus parallèle ou comparatif est nécessaire pour constituer un système de traduction automatique, cet article décrit les étapes de construction d'un tel corpus ainsi que les outils utilisés.

1. Introduction

Le défi fréquemment soulevé dans le domaine des sciences humaines et plus particulièrement des sciences du langage est de briser les barrières linguistiques en créant un espace des échanges socioculturels, politiques, économiques entre les différentes langues qui existent dans le monde. L'un des outils permettant cet objectif est la traduction automatique des langues.

La première partie tente de cerner les différentes méthodes et algorithmes permettant la traduction automatique. Alors que la deuxième partie décrit les principales méthodes de constitution du corpus à savoir : l'extraction des données brutes, l'alignement des documents, et l'alignement des phrases.

2. Les approches de la traduction automatique

2.1. Approches Classiques

Le « triangle de Vauquois » proposé par [Vauquois 1968] pour la traduction décrit les différentes architectures linguistiques possibles d'un système de TA.

Chaque chemin dans le triangle correspond à une architecture linguistique.

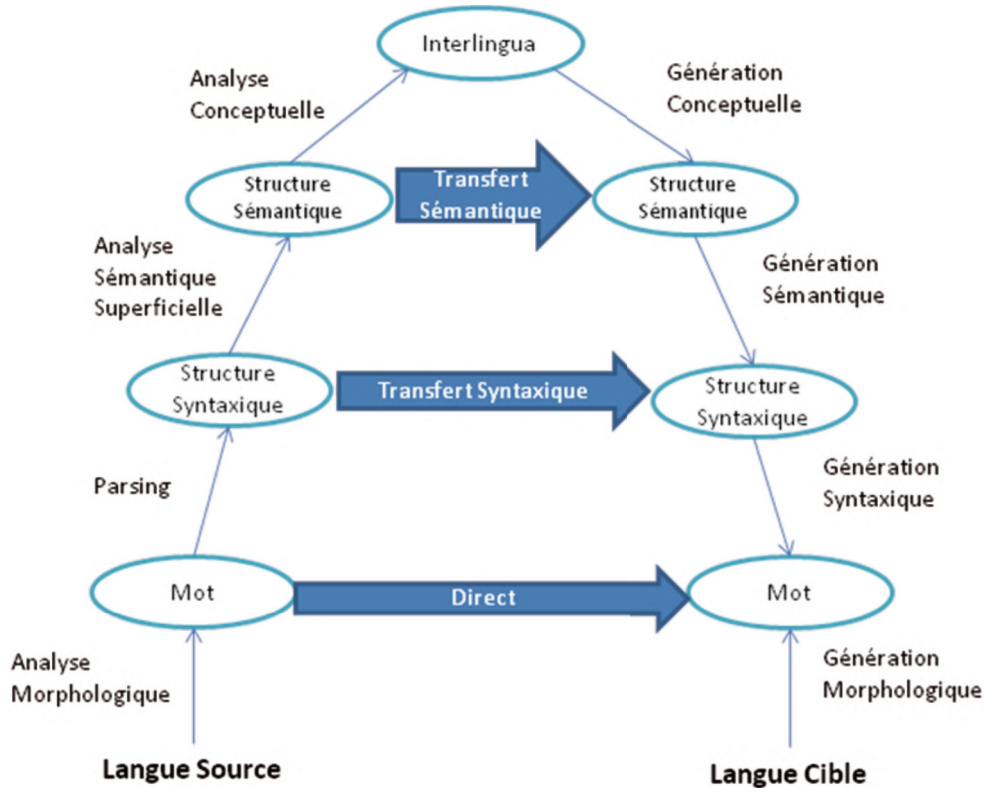


Figure 1 – Triangle de Vauquois

2.1.1. Approche par traduction directe

La phrase source est segmentée par mots et le système exécute la traduction mot à mot sans représentation intermédiaire. Elle n'est pas en mesure de gérer les ré ordonnancements de langue distance puisqu'elle ignore la structure syntaxique.

Le dictionnaire bilingue nécessaire à l'approche directe sera très langue et très difficile à construire

2.1.2. Approche par transfert

En plus de traduire les mots (transfert lexical) on aura besoins de traduire les arbres syntaxiques (transfert syntaxique).

Nous citons l'exemple suivant pour illustration :

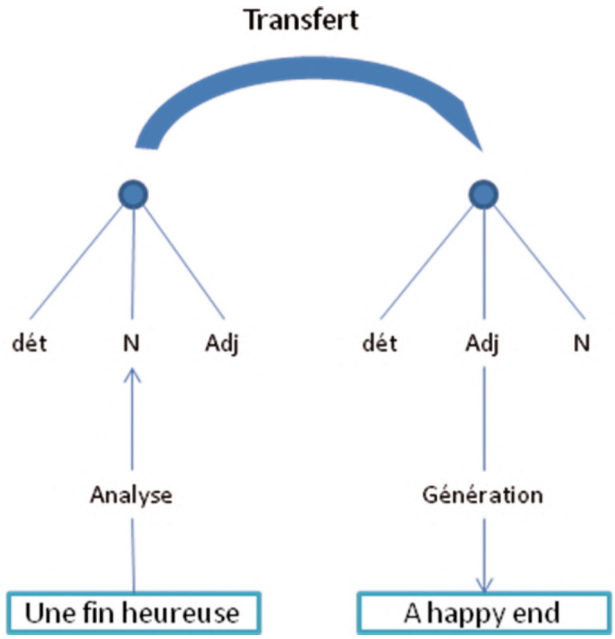


Figure 2 – Arbre syntaxique

Des nombreux systèmes de TA fondés sur cette approche sont disponibles aujourd'hui, tels que Systran, utilisé par Babelfish, Reverso, The Translator (Toshiba), etc.

2.1.3. Approche par interlingua

L'idée est de faire passer l'analyse vers une représentation indépendante de la langue.

UNL :(Universal Networking Language) est un exemple de langage pivot [Uchida 2001] en cours de développement.

2.2. Approches computationnelles

2.2.1. Méthodes expertes

Les méthodes expertes font appel à la programmation basée sur des modèles de calcul abstraits donnant lieu à des langages spécialisés pour la programmation linguistique.

Nous citons ci-après quelques exemples :

- LFG (Lexical-Functional Grammars)
- HPSC (head driven phrase structure grammar)
- TAG (tree adjoining grammar)

2.2.2. Méthodes empiriques

On distingue deux types :

- La TA fondée sur les exemples (EBMT, Example-Based Machine Translation).
- La TA probabiliste (SMT, Statistical Machine Translation).

L'approche de TA fondée sur les exemples a été proposée par [Nagao 1984]

L'idée de base de cette approche est de réutiliser des exemples de traductions existantes comme base pour la nouvelle traduction.

Ci-après les principales étapes de cette méthode :

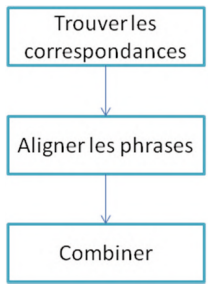


Figure 3 – Algorithme

Plusieurs travaux ont utilisé les méthodes empiriques pour déduire l'équivalent des phrases à traduire dans les langues cibles, nous citons le cas de la traduction anglais – japonais extraite depuis les travaux de [Sato 1990].

Une déduction simple pour trouver la meilleure traduction d'un mot basée sur la similitude sémantique est pris depuis les travaux de [Nagao 1984]

2.3. Approches Probabilistes

2.3.1. Introduction

Se basent sur un modèle probabiliste entre la phrase source qu'on note ici F et la phrase cible notée E

On note :

- $F = f_1, f_2, \ldots\ldots f_m$ une phrase dans la langue source
- $E = e_1, e_2, \ldots\ldots e_l$ une phrase en langue cible

L'approche probabiliste formule le problème de traduction comme suit :

$$\hat{E} = ar\ gmax_E\ P(E|F)$$

$$= ar\ gmax_E \frac{P(E|F)\ P(E)}{P(F)}$$

$$= ar\ gmax_E\ P(E|F)\ P(E)$$

La traduction retournée doit être bien construite en langue cible (modèle de langue P(E)) et doit contenir les information sur la phrase source F (modèle de traduction P(F/E))

A noter que pour traduire de F vers E, on va en faire traduire un modèle de E vers F ce qui est décrit dans le modèle du canal bruité de transmission appliquée largement dans la reconnaissance automatique

Figure 4 – Canal bruité de transmission

L'approche probabiliste se décompose en 3 composants :

- Un Modèle de langue P(E)
- Un modèle de traduction P(F/E)
- Un décodeur qui produit une traduction de E à partir d'une phrase source F

Dans la suite de cet article, On va donc voir comment définir un modèle de traduction et ensuite comment le décoder.

2.3.2. Modèles et techniques d'alignement

Les premiers modèles de traduction statistique étaient basés sur une traduction mot à mot

L'approche phrase-based procède comme suit :

1. Extraire les syntagmes de la phrase E
2. Traduire chaque syntagme de E en un syntagme de F avec une probabilité $\varphi(\bar{f}_i|\bar{e}_i)$ appelée **probabilité de traduction**.
3. Rassembler les syntagmes.

Exemple :

Anglais	|	Listening	is	an art
Français	|	Savoir écouter	est	un art

L'ordre des syntagmes traduis est altéré selon une distribution de **probabilité de distorsion**, définit comme suit :

$$d(a_i - b_{i-1}) = \alpha^{|a_i - b_{i-1} - 1|}$$

Où a_i est la position de début de $\overline{f}_i$ et b_{i-1} est la position de fin de $\overline{f}_{i-1}$

$\alpha > 1$ va favoriser le ré ordonnancement

$\alpha < 1$ va pénaliser le ré ordonnancement

$\alpha = 1$ va les considérer comme équiprobables

Exemple :

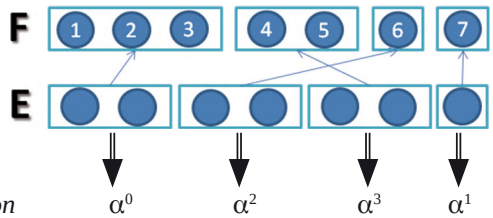
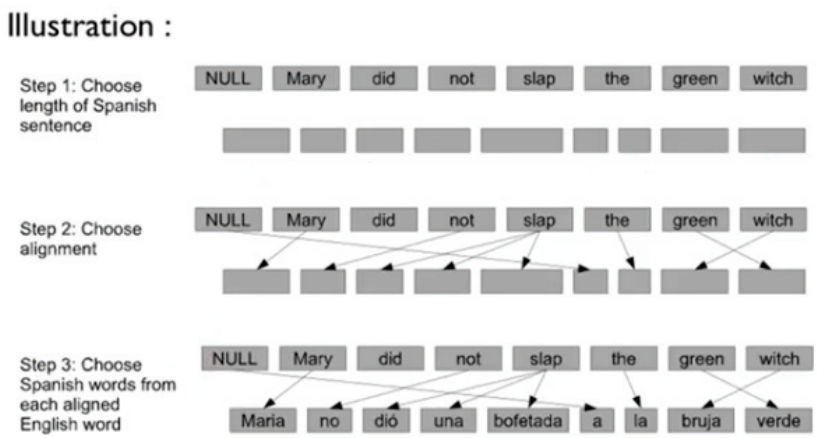


Figure 5 – Exemple de distorsion

La probabilité P(F|E) est alors calculée comme suit :

$$P(F|E) = \prod \varphi(\bar{f}_i|\bar{e}_i) d(a_i - b_{i-1})$$

On va se concentrer par la suite de ce paragraphe au modèle d'alignement simple IBM1 disponible sur le programme GIZA++.



Modèle d'alignement simple IBM1 :

IBM1 suppose qu'une phrase *E* est traduite par *F* comme suit

1. On choisit une taille *J* pour la phrase *F*
2. On choisit un alignement $A = a_1, a_2, \ldots, a_J$ entre la phrase *E* et la phrase *F*
3. Pour chaque position *j* dans la phrase *F*, on assigne un mot f_j, en introduisant le mot e_{aj} auquel il est aligné

2.3.3. Décodage

L'idée est de partir d'une traduction vide (null hypothesis) et de l'étendre en priorisant les traductions partielles prometteuse

Algorithme :

Fonction *Décodage de la pile (phrase source)* **retourner** *(phrase cible)*
Initialiser la pile avec une hypothèse nulle
Boucler
Meilleure hypothèse *h* hors de la pile
Si *h* est une phrase complète. **Retourner** *h*
Pour chaque élargissement possible *h'* de *h*
Assigner un coût a à *h'*
Ajouter *h'* à la pile

3. Les méthodes d'extraction du corpus parallèle

La difficulté qui réside dans la constitution d'un système de TAL pour une langue peu doté comme l'Amazigh, est la constitution du corpus parallèle dont la collecte se compose de trois étapes principales: la collection de données brutes, l'alignement des documents, l'alignement des phrases.

3.1. Extraire des données brutes

Pour constituer le corpus parallèle, [Thi Ngoc Diep Do 2012] propose que Les données soient collectées à partir du web via le protocole http et ftp sous forme de pages web en utilisant des outils tel que GNU wget permettant la collecte de plusieurs pages web entre deux dates précises ensuite un filtrage par langue s'avère nécessaire pour traiter et nettoyer les documents.il a utilisé pour ce cas l'outil Textcat pour le filtrage et l'outil open source CLIPS-Text-Tk pour le traitement et la transformation des pages web en document text. La même démarche a été utilisée par [Thi Ngoc Diep Do 2012] en utilisant l'outil *Mulce*.

3.2. Alignements de document

Il existe plusieurs approches permettant l'alignement des documents à partir d'un corpus de textes, nous pouvons mentionner : le système STAND (Structural Translation Recognition for Acquiring Natural Data) proposé par [Resnik 1998] dont le principe est de considérer que deux document parallèle ont le même url et la même mise en page,Le système PTMiner (ParallelText Miner) de [Chen 2000] utilisé aussi les adresses URLs de documents, avec d'autres critères tels que le rapport de la longueur des documents, la proportion de paires de phrases qui contiennent des éléments d'un dictionnaire bilingue,Le modèle DOM (Document Object Model) présenté par [Shi 2006] pour présenter deux pages web comme des arbres DOM.et des outils existants déjà sur le web tel que Linguee qui est un outil puissant et gratuit permettant l'alignement des documents en ligne.

3.3. Alignements de phrases

Parmi les méthodes les plus connues on cite La méthode se basant sur la longueur des phrases proposé dans les travaux de [Brown 1991] et [Gale 1993].elle considère que la longueur des phrases dans le texte source et celle de leurs traduction dans la langue cibles sont corrélées et La deuxième est fondée sur la langueur des segments présentée dans les recherches de [VERONIS, LANGLAIS 2000].

Références

[Vauquois 1968] Vauquois, B., *A survey of formal grammars and algorithms for recognition and transformation in machine translation*. IFIP Congress-68, Edinburgh, 254–260; reprinted in Ch. Boitet (ed.) Bernard Vauquois et la TAO: *Vingt-cinq Ans de Traduction Automatique – Analectes*, 201–213, Grenoble (1988): Association Champollion. 1968.

[Uchida 2001] Hiroshi Uchida, Meiying Zhu, *The universal networking language beyond machine translation*, UNDL Foundation, Japan, 2001

[Nagao 1984] Makoto Nagao, *A framework of a mechanical translation between Japanese and English by analogy principle, In proceedings of Artificial And Human Intelligence* (A. Elithorn and R. Banerji, editors), Elsevier Science Publishers, B.V, 1984

[Sato 1990] Sato, S., and Nagao, M., *Toward memory-based translation,* In proceedings of COLING 1990

[Resnik 1998] Philip Resnik, Parallel Strands: *a preliminary investigation into mining the web for bilingual text,* In proceedings of AMTA 1998

[Chen 2000] Chen Jiang et Jian-Yun Nie, *Parallel Web text mining for cross language information retrieval,* In proceedings of RIAO 2000

[Shi 2006] Shi L. Niu C., Zhou M., Gao J., *A DOM tree alignment model for mining parallel data from the web*, In proceedings of Computational Linguistics / ACL 2006

[Brown 1991] Brown P.F., Lai J.C. et Mercer R.L, Aligning *sentences in parallel corpora,* In proceedings of ACL 1991

[Gale 1993] W.A. Gale, K.W. Church, *A program for aligning sentences in bilingual corpora,* In proceedings of ACL 1993

[VERONIS, LANGLAIS 2000] VERONIS Jean, LANGLAIS Philippr (2000). *Parallel Text Processing : Alignement and use of translation corpora*. Academic Publishers, pp. 369-388.

[Thi Ngoc Diep Do 2012] Thi Ngoc Diep Do. *Extraction de corpus parallèle pour la traduction automatique depuis et vers une langue peu dotée*. Autre [cs.OH]. Université de Grenoble, 2011. Français. <NNT : 2011GRENM065>. <tel-00680046>

[Ciara R. Wigham, Aurélie Bayle 2012] Ciara R. Wigham, Aurélie Bayle. *Enjeux, outils et méthodologie de constitution de corpus d'apprentissage*. Coldoc 2012 : Traitements de corpus : outils et méthodes, Oct 2012, Paris, France. <edutice-00710698>

TICAM'16, p. 159 - 167

IT Communication as institutional promotion for the technological realizations: Case of the Center for Informatics Studies and Information' Systems and Communication

Mounia Boumediane
Centre des Etudes Informatiques, Systèmes d'Information et de Communication, IRCAM
boumediane@ircam.ma

Abstract

Internet along with groupware tools may well be viewed to be the prime reasons behind the spreading and dissemination of information technology (IT) communication. IT communication, not with standing its use of structured data along different information systems, represents only a very small number of the total volume of the manipulated information. The documentation sent via mail or visited through the web is written in natural language and, thereby, exploits the flexibility and the communicative power of natural languages. With the above in mind, let us address the following questions more thoroughly:

- How can the Royal Institute of Amazigh Culture take advantage of IT communication' potentials?
- What is the nature of the problems that IT communication may confront and how can we solve them?

To answer to those problematic research questions, we have to undergo the realizations made by and within the Royal institute of Amazigh Culture by the center for Informatics Studies, Information Systems and Communication.

1. Introduction

Technologies related to the Internet, mobile are profoundly affecting institutions worldwide at many levels: social, economic and political, particularly in emerging democracies. In the hands of reformers and activists, these tools can overcome resource disparities and entrenched monopolies of power and voice. Citizens, civil and non-governmental organizations, companies, politicians, administrations and large state and private-sector bureaucracies are employing technologies and the Internet to enhance communication, improve access to important information, and increase their efficiency, resulting in strengthened democratic processes and more effective governance. "Recent and predicted future changes in the leadership of many Middle Eastern and North African countries, most notably Morocco and Jordan, and Syria and elsewhere have led media pundits to note the coming of an «Internet

Leadership» in the region. The term alludes to the fact that the new and up and coming leadership in the region is relatively young, educated in the US or Europe, recognizes the key role that technology in general and ICT (information and communication technology) in particular play in national development. Despite the perceived risks from joining the global computer network, non-democratic regimes have not shunned this technology because it promises to be the key to development, prosperity, and influence"[1].

According to UNESCO, almost half of the approximately 7.000 currently spoken languages are expected to become extinct this century. It is estimated that less than 5% of these will be used online or have significant digital presence[2]. Languages which do not have significant digital resources risk extinction from loss of prestige, specific domain usage such as online, and ultimately loss of speakers.

Many language communities, academics and institutions are working to prevent this by developing tools, websites, and resources for their languages. The Royal Institute of Amazigh Culture as an academic and public institution whose main aims are to develop and to promote the amazigh language and culture has sustained and encouraged technological projects which facilitate the learning and acquisition of the Amazigh language for non native speakers whether for adults or for children. In this aspect, the center of informatics studies and information system and communication has developed a series of digital projects. The objectives of this article is to shed light on these projects and to emphasize the potentials of IT communication for and by the Royal Institute of Amazigh Culture in relation to its technological realizations and to project the nature of the problems that IT communication recognized while dealing with Amazighe language.

2. Identity card of Royal Institute of Amazigh Culture (RIAC)

Royal Institute of Amazigh Culture, whose abbreviation in French is IRCAM, is an academic institution created by Royal Dahir. It is endowed with a full legal competence and financial autonomy.

2.1.Missions of the Royal Institute of Amazigh Culture

As far as the missions of IRCAM are concerned, the principals ones are:

- To preserve and to promote amazighe culture.
- To renforce the place of amazighe culture in the social and cultural and media fields.
- To develop the cooperation with other national and international institutions and organisations working in the domain.

1 Mason, Moya K. (2012), Potential of Information and Communication Technologies (ICTs) for Developing Countries, The High-level Forum on City Information in the Asia-Pacific Region. In http://www.moyak.com/papers/ngo-icts.html. Internet Usage in the Middle East: Some Political and Social Implications. Consulted in 04 june 2016.

2 Kornai, A. (2013). Digital language death. PloS one; 8(10): e77056.

To fulfill these missions, the Royal Institute of Amazigh Culture depends on its structural research and administration. The research structure is composed of seven centers and each one contains two units of research as mentioned in below:

Figure1: Centers of research and its units of research

For the administrative structure, is composed of departments and services embeded with tremendous young human resources of different disciplines referred to technical's specialty and administrative competencies headed by the Rector and the General Secretary.

3. Advantages of IT Communication

IT communication for and by the administrative and academic structure of the Royal Institut of Amazigh Culture offers many advantages. In fact, as new communication methods, IT enhances the contact with others for less money and over greater distances than ever before. Technologies such as texting, instant messaging and video conferencing, free applications allow users to communicate instantaneously with people across the world for a nominal fee. A "concept which may have seems ludicrous before the advent of computers. In addition, text-based computer communication can give those with speech or social problems a level playing field to communicate with their peers."[3]

Bearing in mind those advantages, we can ask the following question how can the Royal Institute of Amazigh Culture take advantage of the potentials of IT communication?

3 Andy Walton, (2012), Advantages & Disadvantages of Information & Communication Technology, In http://smallbusiness.chron.com/advantages-disadvantages-information-communication-technology-66948.html, consulted in 1 June 2016.

The NTIC have permeated in many daily activities, whether shopping, sending email and recording voice messages on a cell phone, booking travels on line. It seems as no surprise that NTIC have co-opted by the field of education. Permeation of technology is seen everywhere. In Morocco, and to take the example of the Royal Institute of The Amazighe Culture, as an academic institution, which main objective is the promotion and developpement of the amazighe language and culture, has integrated an educative web site within its external web site since its creation. This educative web site is an interactive mean of learning amazighe language for non native speakers wherever they are. The IRCAMIEN educative web site is composed of three columns as shown in the figure below. We find column of Amazigh School, learning tifinaghe, pedagogic support. The amazigh school CD is composed of three components: library, CD-rom, contact.

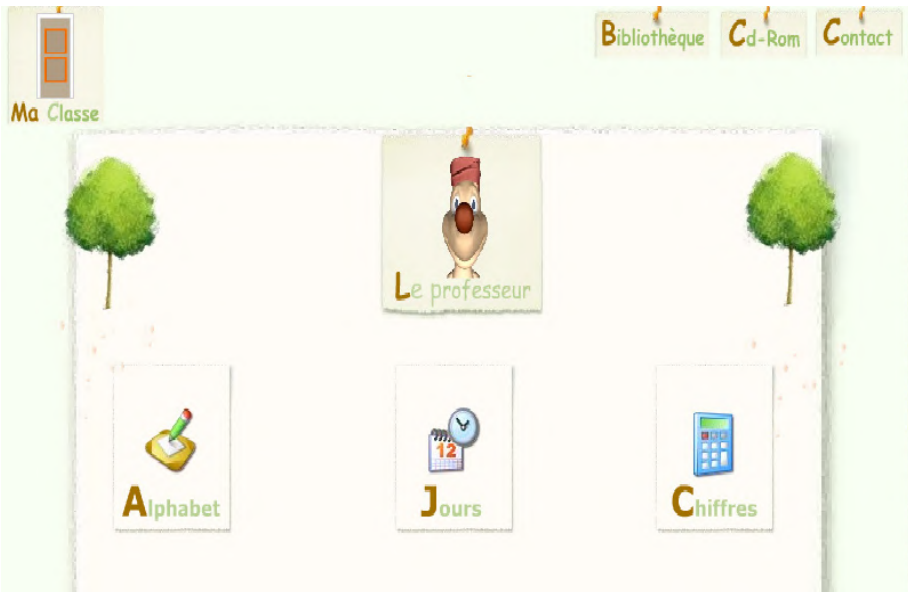


Figure2: extract of the school Amazighe developed since 2004.In www.ircam.ma.

In fact, The Center of Informatics Studies and Informational System and Communication is the center who has the mission to develop tools and applications to enhance and supply the teaching and learning of amazighe language; whether for adults or children with appropriate ones, with adequate technological means in collaboration with the other centers whose mean missions are the didactic and pedagogy. In this aspect, the center for informatics studies and informational system and communication had and is still producing systems and application via cd and mobile that empowers amazighe language learning.

In fact, the tifinagh as graph and characters have been introduced to the world web. The majority of the navigators can support tifinagh, the figure below indicates the navigators which support the tifinagh:

Figure 3: the navigators supporting tifinagh

This accessibility to the virtual world has enhanced the amazigh language and culture and has facilitate the developers to adventure in the production and realization of tools and applications intended to promote the learning of the amzighe language.

In this aspect, the center of informatics studies, system of information and communication has realized and led IT projects that aim to promote the Amazigh language and culture.

3. Technological Realizations of the center of informatics studies and system of information and communication

The center of informatics studies and system of information and communication had produced many applications, resources in order to facilitate the accompaniment of the learning and teaching of amazigh language. The following chart lists some of them:

Figure 4 : Source : https://play.google.com/store/apps/details?id=com.ircam.vocabetlettres

My first words in amazigh named in Tamazight ⵙⵍⵓⵣⵉⵙⵎⵛⵍⵓⵙ is an educational application for children aged between 2 to 12 years.

The main objectif of this application is to motivate and support the learning of the Amazigh language for children. It includes the learning of vocabulary amazigh axed on six themes namely, fruits, vegetables, animals, shapes, colors and human body and calculation. Also the application includes a writing module based on the writing of the 33 the amazigh characters.

Figure 5 : Source : http://tal.ircam.ma/talam/lexam.php

Electronic lexicon Amazigh mobile (LEXAM in French abbreviation) is the first mobile application android with a standardized Moroccan Amazigh lexicon covering areas of everyday modern life referring to: media , administration, art , environment, civility, law and justice, education, and many others.

The current version contains over 4000 Amazigh entries.

Figure 6 : Source : http://tal.ircam.ma/tamawalt/cat.php

ⵜⴰⵎⴰⵡⴰⵍⵜ is a multilingual dictionary which aims to introduce children to learning of the Amazigh language.

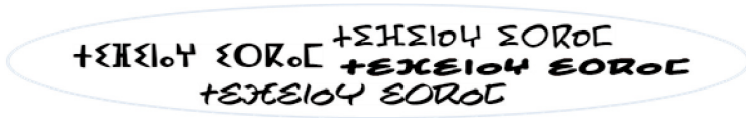
The site of this dictionary is dedicated to teaching the amazigh language. It is intended for younger children as for older ones, from kindergarten to the middle school to the primary one. The site consists of six sections (homepage, categories, alphabet, games including memory games, letters and numbers storing and classification games, search, contact).

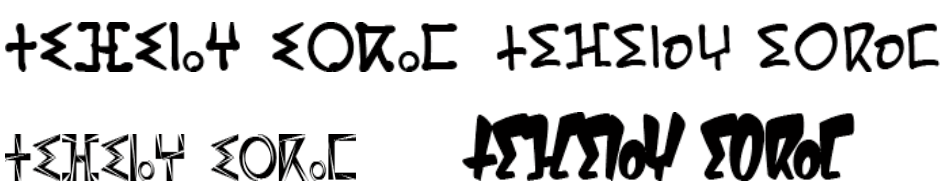
Figure 7 : Source : http://www.ircam.ma/?q=ecole_amazighe

A set of educational CDs designed to consolidate learning the Amazigh language. Meanwhile, it promotes Amazigh through knowledge sharing.

Figure 8 : Source : http://tal.ircam.ma/tamawalt/cat.php

Portal Automatic Processing of the Amazigh language (TALAM) is a project led by the Centre of Computer Studies of Information and Communication Systems (CEISIC), whose aim is to provide a portal within a set of linguistic resources digital and automatic processing tools available to the Amazigh documentation and language. It fonctions as a technical platform and scientific equipment for the written and oral Amazigh language and aims to facilitate the exploitation and transfer of resources and tools developed within the IRCAM to public laboratories.





Many Police Tifinagh characters has been developed which fortify the stylistic and aesthetic merits of the Amazigh writing.

4. Nature of the problems in relation to IT amazigh Communication

To undergo this problematic statement whether the amazigh language constitute an obstacle for developers when dealing with Amazigh language in IT projects, we have elaborated a sociolinguistic research of qualitative nature based on an indirect structured interview [4]addressed to researchers in Rabat, Italy, France. Whoever, only researchers in Rabat have answered to the quest. The sample of the Rabat, which is representative and non exhaustive, is composed mainly of the researchers of the Mohamadia Engineering School. The conclusions of the response of researchers are based on the following responses:

1. The Amazigh language belongs to the Hamito-Semitic/"Afro-Asiatic" languages, with rich templatic morphology. In linguistic terms, the language is characterized by the proliferation of dialects due to historical and geographical factors. It is spoken in Morocco, Algeria, Tunisia, Libya, and Siwa (an Egyptian Oasis); it is also spoken by many other communities in parts of Niger and Mali. It is used by tens of millions of people in North Africa mainly for oral communication and has been introduced in mass media and in the educational system in collaboration with several ministries in Morocco. In Morocco, for instance, one may distinguish three major dialects: Tarifit in theNorth, Tamazight in the center and Tashlhit in the southern parts of the country. It is a composite of dialects which none have been considered the national standard,
2. It is a language with its own script. The official graphic system for writing Amazigh is Tifinagh. It does not have capitalization in its script and it is written from left to right. IRCAM kept only pertinent phonemes for Tamazight, so the number of the alphabetical phonetic entities is 33 (consisting of: 27 consonants, 2 semi-consonants and 4 vowels), however Unicode codes only 31 letters plus a modifier letter to form the two phonetic entities: ⵅⵯ(g^w) and ⴽⵯ(k^w). The whole range of Tifinagh letters is subdivided into four subsets: the basic letters used by IRCAM, an extended set used

4 Indirected structured interview is self elaboration composed of a direct question but administrated indirectly by means of informant.

also by IRCAM, other neo-tifinaghe letters in use and some attested modern Touareg letters. The total number of Tifinagh letters after the two amendments reaches 59 characters, and are occupying 2D30-2D7F plage in Unicode;

3. Ambiguity (like the other languages); the same surface form might have many meanings or POS(Part Of Speech) tag depending on how it has been used in the sentence;
4. Scarcity of tools and resources: like most of the languages that has only recently started being investigated for the NLP tasks, Amazigh lacks of annotated corpora and processing tools and resources.

5. Conclusion

Notwithstanding the fact that the computerization of Amazigh has become a determining factor in its use and its survival but the communication up on this computerization is much more meaningful. For this reason, the Amazigh language needs a dynamic company in which IT communication can play a key role in language processing, where they can be an important vehicle for communication for the metalanguage and also between and within language communities. Otherwise, not taking advantage of IT communication, it can be assumed to be an aggravating factor in the marginalization of a language.

References

HAYES, M., (2006). ICT in the Early Years. Edited by Whitebread David. Publisher Open University Press, England.

Kornai, A. (2013). Digital language death. PloS one; 8(10): e77056.

Kozma, R. (2003), Technology, Innovation, and Educational Change: A Global Perspective: A Report of the Second Information Technology in Education Study, Module 2. Editor: International Society for Technology in Education.

Mason, Moya K. (2012), Potential of Information and Communication Technologies (ICTs) for Developing Countries, The High-level Forum on City Information in the Asia-Pacific Region.

World bank, (2015), ICT for Greater Development Impact Information & Communication Technology.

http://siteresources.worldbank.org/EXTINFORMATIONANDCOMMUNICATIONANDTECHNOLOGIES/Resources/WBG_ICT_Strategy-2012.pdf. Consulted in 04 July 2016.

www.ircam.ma.

La correction automatique d'orthographe dans la langue arabe

Mohammed Nejja[1] **Abdellah Yousfi[2]**
[1] Équipe Telecom et systèmes embarqués, ENSIAS, Université Mohammed V - Rabat- Maroc
mohammed.nejja@gmail.com
[2] Équipe ERADIASS - FSJES, Université Mohammed V - Rabat- Maroc
yousfi240ma@yahoo.fr

Résumé

Ce papier porte sur la correction automatique des erreurs orthographiques dans la langue arabe. La présente étude présente un mécanisme à utiliser pour résoudre le problème proposé par le correcteur d'orthographe qui réside dans l'insuffisance du lexique utilisé. Dans ce contexte, nous avons traité deux axes : La correction automatique d'orthographe en utilisant le schème de surface et la gestion de vocabulaire.

1. Introduction

Le système de correction automatique des textes est l'un des plus anciennes applications des techniques de traitement automatique des langues (Mitton, 2010). La correction orthographique consiste à trouver le mot le plus proche au celui erroné en se basant généralement sur un dictionnaire contenant l'ensemble des mots de la langue donnée.

Les premières techniques de la correction automatique d'orthographe consistent à chercher le mot dans le dictionnaire, elles prospectent simplement si la chaîne en entrée apparaît ou non dans la liste des mots valides. Si la chaîne est absente du dictionnaire alors elle est dite erronée et le système propose à l'utilisateur des candidats les plus proches au celui erroné.

Aujourd'hui, plusieurs techniques sont disponible pour détecter et corriger les fautes d'orthographe, notamment les systèmes s'appuyant sur des automates à états finis (pondérés) (Ringlstetter *et al.* , 2006), Les méthodes adoptant le modèle du canal bruité (Kernighan *et al.* , 1990; Brillet Moore, 2000; KolaketResnik, 2002), des systèmes basés sur le principe d'utilisation des HMM, Les techniques basées sur le modèle n-gramme (Ukkonen,1992), la distance d'Edition…

Dans cet article, nous nous sommes intéressé à la correction automatique d'orthographe dans la langue arabe via la technique de la distance d'édition en exploitant les informations morphologiques des mots. Ces techniques dont nous citons visent à réduire la taille du dictionnaire utilisé ainsi que le nombre des mots à comparer afin d'assurer la performance de système de correction automatique d'orthographe.

2. La distance de levenshtien

La distance de Levenshtein (Levenshtein, 1966) est une distance mathématique donnant une mesure de la similarité entre deux chaînes de caractères. Elle est égale au nombre minimal de caractères qu'il faut supprimer, insérer ou remplacer pour passer d'une chaîne à l'autre.

L'algorithme de levenshtein utilise une matrice de (N + 1) * (P + 1) « avec N et P les longueurs des chaînes de caractères à comparer T, S». la distance entre les chaines T et S est définie par :

$$M(i, j) = \text{Min} \begin{cases} M(i - 1, j) + 1 \\ M(i, j - 1) + 1 \\ M(i - 1, j - 1) + \text{Cout}\ (i - 1, j - 1) \end{cases} \quad (1)$$

Avec

$$\text{Cout}\ (i, j) = \begin{cases} 0 & \text{si} \quad T(i) = S(j) \\ 1 & \text{si} \quad T(i) \neq S(j) \end{cases} \quad (2)$$

3. Les schèmes de surface

Le schème arabe permet essentiellement de déterminer la structure de la plupart des mots (les noms, les verbes conjugués, etc.). Les schèmes sont des déclinaisons du mot «فعل » qui sont obtenus en utilisant des diacritiques ou en y ajoutant des affixes.

Le schème de 'يَكْتُبُونَ' est يَفْعَلُونَ. Les lettres « ف،ع،ل » remplacent les lettres de la racine de يَكْتُبُونَ, et le schème de « نَالَ » est « فَعَلَ».

Ce type de schème ne peut pas présenter les variations morphologiques du mot (par exemple le nom قَائِلٌ du verbe (قَال). C'est pour cela que M. Yousfi a proposé un schème adapté appelé schème de surface (Yousfi, 2010).

Exemple :

La conjugaison du mot « رَعَى» au participe actif à la 1ère personne du singulier, est « رَاعٍ », alors le schème de surface de la racine « رَعَى » est « فَعَى» et « فَاعٍ » est le schème de surface de « رَاعٍ ».

Le schème de surface de « آجِرٌ» est «آفِعٌ « et de «آجِرَاتٌ» est «آفِعَاتٌ».

4. La correction morphologique par des schèmes de surface dans l'algorithme de levenshtein

4.1. Objectif

Notre objectif consiste à remédier aux problèmes proposés par les systèmes de la correction automatique d'orthographe, et donc fournir un correcteur pertinent. Parmi les problèmes étudiés, nous citons l'insuffisance du lexique utilisé, en fait le principe de détection de l'erreur consiste à se servir d'un dictionnaire de taille très large contenant tous les mots de la langue dans leurs différentes formes, ce qui est généralement difficile à construire. Pour remédier à ce problème, nous avons proposé une solution en exploitant la propriété de la dérivation des mots arabe. En fait, la plupart des mots arabe sont des déclinaisons (mot dérivé) du mot فعل en ajoutant des affixes et/ou diacritique.

Exemple :

Racine	Schème	Mot dérivé
كتب	مفعول	مكتوب
لعب	ستفعلون	ستلعبون
شرب	فاعل	شارب

4.2. La solution proposée

La méthode que nous avons proposé (Nejja et Yousfi, 2014 ; Nejja et Yousfi, 2015) est basée sur deux dictionnaires, un contient l'ensemble des racines, l'autre contient l'ensemble des schèmes. Donc à partir d'une base de données contenant 10000 racines et 10000 schèmes, nous pouvons traiter (vérifier et corriger) jusqu'à 100000000 mots.

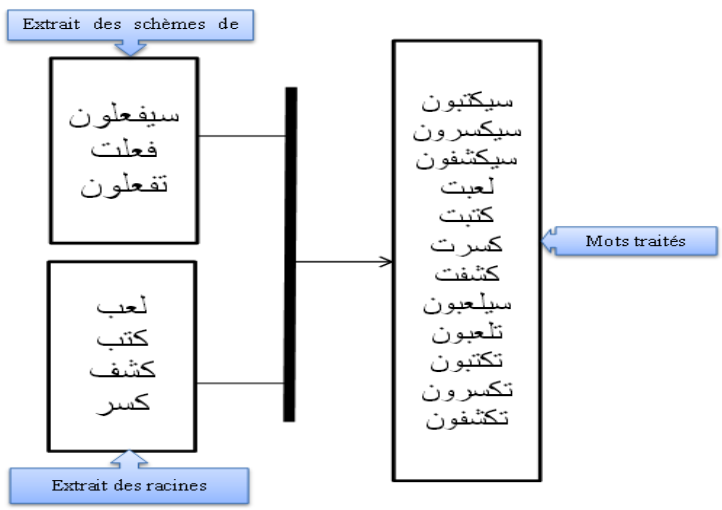


Figure. 1. Extrait des mots traités

Pour mettre en œuvre cette idée, nous avons élaboré des approches basées dans un premier temps sur l'identification du schème de surface le plus proche au mot erroné. Pour ce faire, nous avons adapté la distance de levenshtein à la langue arabe pour extraire les schèmes de surface les plus correct.

La distance de levenshtein adaptée à la sélection du schème de surface (noté A) et le mot en entrée (noté w) est définie par :

$$M(k, p) = \text{Min} \begin{cases} M(k - 1, p) + 1 \\ M(k, p - 1) + 1 \\ M(k - 1, p - 1) + \text{Cout}\ (k - 1, t - 1) \end{cases} \quad (1)$$

Avec

$$\text{Cout}\ (k, p) = \begin{cases} 1 & \text{si}\ \ A(k) \neq w(p) \text{ et } A(k) \notin \{'ف', 'ع', 'ل'\} \\ 0 & \text{si}\ \ A(k) = w(p) \text{ et } A(k) \in \beta\{'ف', 'ع', 'ل'\} \end{cases} \quad (3)$$

Exemple

Soit : le mot ستلعبون et le schème de surface ستفعلون donc la distance de levenshtein adapté est 0 :

		س	ت	ل	ع	ب	و	ن
	0	1	2	3	4	5	6	7
س	1	0	1	2	3	4	5	6
ت	2	1	0	1	2	3	4	5
ف	3	2	1	0	1	2	3	4
ع	4	3	2	1	0	1	2	3
ل	5	4	3	2	1	0	1	2
و	6	5	4	3	2	1	0	1
ن	7	6	5	4	3	2	1	0

Figure. 2. La matrice de calcul de la distance de levenshtein adaptée

Cependant, cette approche présente aussi un inconvénient qui réside dans le fait que le schème de surface désiré peut être parfois situé en dernière position. Cela est due au grand nombre des solutions proposées ayant la même distance.

Exemple :

Pour le mot erroné ستتبون nous avons obtenu le résultat suivant :

Edit Distance	Surface Patterns
1	ستفعلون
1	سيفعلون
3	نفاعلوا
…	…

Figure. 3. Extrait des schèmes de surface proposés pour le mot ستتبون *par la formule 3*

Pour remédier à cette lacune, nous avons amélioré la formule utilisée pour la sélection de schème de surface en partant de l'hypothèse que chaque caractère peut être un caractère de la racine ou bien un caractère de schème de surface :

$$M(k, p) = \text{Min} \begin{cases} M(k - 1, p) + 1 \\ M(k, p - 1) + 1 \\ M(k - 1, p - 1) + \text{Cout}\ (k - 1, t - 1) \end{cases} \quad (1)$$

Avec

$$\text{Cout}\ (k, p) = \begin{cases} 1 & \text{si } A(k) \neq B(p) \text{ et } A(k) \notin \{\text{'ف'}, \text{'ع'}, \text{'ل'}\} \\ 0 & \text{si } A(k) \in \{\text{'ف'}, \text{'ع'}, \text{'ل'}\} \\ 0{,}2 & \text{si } A(k) = B(p) \text{ et } A(k) \notin \{\text{'فع'}, \text{'ل'}\} \text{ et } p = k \\ 0{,}5 & \text{si } A(k) = B(p) \text{ et } A(k) \notin \{\text{'ف'}, \text{'ع'}, \text{'ل'}\} \text{ et } p \neq k \end{cases} \quad (4)$$

La modification que nous avons introduit dans la formule nous a permis d'augmenter la précision de la sélection de schème de surface. Pour le même exemple, nous avons obtenu le résultat suivant :

Edit Distance	Surface Patterns
1.9	ستفعلون
2.2	سيفعلون
2.5	نفاعلوا
…	…

Figure. 4. Extrait des schèmes de surface proposés pour le mot ستتبون *par la formule 4*

Une fois le schème de surface est identifié, nous créons un ensemble des candidats potentiels par la dérivation de toutes les racines en utilisant le schème de surface identifié. Enfin nous comparons le mot en entrée avec ces candidats potentiels en utilisant la distance de levenshtein.

Exemple :

soit

- Le mot erroné ستتبون
- Le schème de surface le plus proche : ستفعلون
- Liste des racines [ذهب,كتب, خرج…]
- Donc les candidats potentiels sont [… ستخرجون, ستكتبون, ستذهبون]

En comparant ces candidats potentiels avec le mot erroné, nous proposons le mot ستكتبون comme une solution, car il s'agit de la solution la plus proche lexicalement au mot erroné.

5. La gestion du corpus et la morphologie dans la correction automatique d'orthographe

La réduction du temps de correction nécessaire pour détecter et corriger les fautes d'orthographe dans un texte est vue sous 1'angle de l'optimisation du nombre d'accès au vocabulaire, cette optimisation est restée, depuis longtemps, un défi.

La solution que nous avons proposée (Nejja et Yousfi, 2016) est fondée sur l'utilisation d'un corpus noté C constitué par un ensemble de dictionnaires noté au lieu d'un seul. Chacun d'eux se compose de mots dérivés uniques et non redondants d'un dictionnaire à l'autre, étant donné que le vocabulaire de la langue arabe est essentiellement construit à partir des racines.

$$C = Dict_{R_1} \cup Dict_{R_2} \cup ... \cup Dict_{R_n} \qquad (5)$$

La segmentation du corpus vise à réduire le nombre des mots à comparer en limitant l'accès uniquement au dictionnaire avec la plus grande probabilité de contenir le mot désiré et respectant la règle ci-dessous :

$$\{ (S_i\ R_i \cap w_{(err)} \text{ est non vide @ alors nous prenderons en considération ce dictionnaire @@ } S_i\ R_i \cap w_{(err)} \text{ est vide @ alors nous éliminons ce dictionnaire} \quad (6)$$

Avec R_i la racine des mots qui font partie du dictionnaire i.

La précédente règle nous évitera de faire des calculs inutiles et d'accéder à l'ensemble des données. Seuls les dictionnaires répondant à la règle seront consultés ce qui résulte en un gain en temps de correction.

Pour une bonne illustration, considérons l'exemple suivant:

Soit le corpus rétrécit C contenant les 5 dictionnaires suivants:

- dict 1 contenant les mots dérivés de la racine شرب
- dict 2 contenant les mots dérivés de la racine حذف
- dict 3 contenant les mots dérivés de la racine لعب
- dict 4 contenant les mots dérivés de la racine كنز
- dict 5 contenant les mots dérivés de la racine كره

Supposons que chaque dictionnaire contient 10000 mots dérivés et sachant que le nombre des mots dérivés d'une racine dans la langue arabe est largement supérieur à celui de l'exemple (10000).

Pour le mot erroné «ملغب» et en appliquant la règle au-dessus, nous avons accès uniquement aux dictionnaires 1 et 3, car pour les autres dictionnaires $R_i\ W_{err}$ est vide. Ainsi, le temps d'exécution se voit réduit vu que nous gagnons le temps d'accès, de parcours et de comparaison avec les (3 * 10000) mots des dictionnaires restants.

Le corpus implémenté est organisé selon l'architecture suivante:

Partition 1 regroupe tous les mots non assimilés par le système morphologique dérivationnel de l'arabe. Cette partition contient les mots outils à savoir les prépositions et les pronoms, etc..ainsi que les mots vides et les mots propres tels que les noms des villes, les noms des objets, etc....

Partition 2 Cette partition est construite par tous les mots arabes assimilés par le système morphologique dérivationnel de l'arabe, elle contient une centaine de sous dictionnaires possédant chacun son propre vocabulaire à base de mots dérivés [à partir d'une racine] uniques et non redondants d'un dictionnaire à l'autre.

Pour vérifier et corriger un mot en entrée, nous procédons comme suite :

Nous avons commencé la vérification et correction par la partition 1 en utilisant l'algorithme de levenshtein. Si le mot en entrée apparaît dans la liste des mots outils ou bien les mots vides (distance de levenshtein égale à 0) nous mettons fin à la vérification (le mot fait partie de la langue). Si la chaîne est absente du dictionnaire (aucune distance de levenshtein n'est égale à 0) alors nous gardons les solutions les plus proche au mot en entrée puis nous sélectionnons les sous dictionnaires appartenant à la partition 2 et respectant la règle 5. En utilisant l'algorithme de levenshtein, nous comparons le mot en entrée avec tous les mots dérivés appartenant aux sous dictionnaires sélectionnés. Si le mot existe dans un sous dictionnaire (distance de levensthein égale à 0), nous mettons fin à la vérification. Sinon, nous rassemblons les mots proposés comme solutions au mot en entrée avec les solutions que nous avons déjà gardé pour les organiser selon la distance de lenvenshtein et enfin nous proposons un ensemble des mots les plus proches au mot en entrée.

6. Conclusion

Dans ce travail, nous avons proposé deux techniques utilisées de vérification et éventuellement la correction des mots de la langue arabes sur le plan lexical. La première se base sur l'identification des schèmes de surface les plus proche au mot en entrée. Dans ce contexte, nous avons adapté l'algorithme de levenshtein à la langue arabe puis corriger le mot en entré selon le schème de surface identifié. Pour la deuxième technique, nous avons proposé une gestion de dictionnaire qui nous a permis de limiter l'accès à toutes les données du dictionnaire pour réduire le nombre des mots à comparer en évitant les comparaisons inutiles.

Références

Ringlstetter C., Schulz K. U., Mihov S. and Louka K. (2006). The same is not the same— postcorrection of alphabet confusion errors in mixed-alphabet ocr recognition. *Proceedings of the 8th International Conference on Document Analysis and Recognition (ICDAR'05)*, pp. 406–410.

Kernighan M., Church K. and W.A. G. (1990). A spelling correction program based on a noisy channel model. *Proceedings of the 13th conference on Computational linguistics,* p. 205–210.

Brill E., Moore R. C. (2000). An Improved Error Model for Noisy Channel Spelling Correction. *Proceedings of the 38th Annual Meeting on Association for Computational Linguistics* (ACL'00), pp. 286–293.

Kolak O., Resnik P. (2002). OCR error correction using a noisy channel model.*Proceedings of the Second International Conference on Human Language Technology Research (HLT'02)*. pp. 257–262.

Brucq. D., EI Youbi. A. (1996). Representation de chaines de caracteres par des chaines induites de Markov.Actes RFIA. pp. 651-658.

Ukkonen, E. (1992). Approximate string matching with q-grams and maximal matches. *Theoretical Computer Science*. pp. 191-211.

Levenshtein.V. (1965). Binary codes capable of correcting deletions, insertions, and reversals. *Cybernetics and Control Theory*, pp. 707–710.

Yousfi, A. (2010). The morphological analysis of Arabic verbs by using the surface patterns. *International Journal of Computer Science Issues*, Vol. 7.

Nejja M., Yousfi A. (2014). Correction of the Arabic derived words using surface patterns. *Workshop on Codes, Cryptography and Communication Systems.*

Nejja M., Yousfi A. (2015). A lightweight system for correction of Arabic derived words. *MediterraneanConference on Information & Communication Technologies.*

نحو معجم حاسوبي للمتلازمات اللفظية في اللغة العربية المعاصرة

أيمن الغامدي [1] **و إريك آتويل** [2]
كلية الحاسب، جامعة ليدز بالمملكة المتحدة
[1]scaaa@leeds.ac.uk
[2]e.s.atwell@leeds.ac.uk

ملخص البحث

تهدف الدراسة إلى بناء قائمة للمتلازمات اللفظية كخطوة أولى لمشروع أكبر يهدف لبناء معجم حاسوبي شامل لهذه العبارات في اللغة العربية المعاصرة مبني على عدد من أكبر المدونات العربية المتاحة، وذلك ليكون مصدراً لغويا مفيداً في مجال المعالجة الحاسوبية للغة العربية، وكذلك في تعلم وتعليم اللغة العربية وخاصة لغير الناطقين بها. في هذه الورقة تقرير عن تجربة حاسوبية تم عملها لاستخراج عدد من المتلازمات اللفظية، والتي يمكن تصنيفها إلى عدة فئات بناء على المستويات اللغوية المعروفة صوتيا وصرفيا ونحويا ودلاليا. وفي التجربة الحالية استخرجت قائمة تتكون من أكثر 600 عبارة عن طريق تطبيق نموذج حاسوبي لمعالجة المدونات العربية واستخراج هذه العبارات بطريقة شبه آلية ويتكون النموذج من ثلاث مراحل رئيسة مصحوبة بعدد من المهام الفرعية، وقد اعتمد الباحثان في هذه التجربة على الجمع بين المنهج الحاسوبي الآلي والمنهج الوصفي المبني على التدقيق اليدوي لضمان الجودة النهائية لقائمة المتلازمات اللفظية.

1- مقدمة

تعتبر ظاهرة المتلازمات اللفظية في اللغات الإنسانية من الظواهر التي لفتت نظر الباحثين في مختلف التخصصات التي لها علاقة بدراسة اللغة، فنجد كثيراً من الأبحاث على سبيل المثال في تخصصات: (اللسانيات، علم النفس اللغوي، تعليم تعلم اللغات ، معالجة اللغات الطبيعية ،الذكاء الاصطناعي... وغيرها) تناولت ظاهرة المتلازمات اللفظية من مختلف الزوايا العلمية، وذلك لأهميتها ودورها الحيوي في الوصول إلى فهم حقيقي شامل للغات البشرية والقدرة على تطويع الآلة (أو الحاسب) وتزويدها بمعرفة أوسع لهذه الظاهرة حتى تتعزز قدرتها على تحليل وفهم اللغات البشرية.

على سبيل المثال، في علوم اللسانيات واللسانيات التطبيقية خاصة نجد كثيرا من الأبحاث التي تخصصت في فهم وتحليل هذه الظاهرة وشرح دورها المحوري في عملية فهم اللغة وخاصة في ما يتعلق بالمعالجة الحاسوبية للغة، وكذلك تعلم وتعليم اللغة للناطقين بغيرها، فقد أكدت عدد من الأبحاث اللسانية على أن المعجم العقلي الذي يستعمله الأنسان العادي في لغة التواصل اليومية يتكون من مجموعة من المتلازمات اللفظية، وذلك بدلا من المفهوم السائد قديما والذي يدعي أصحابه أن المعرفة اللغوية ممثلة بكلمات مفردة منعزلة في العقل البشري (Pawley & Syder, 1983; Sinclair, 1987; Kjellmer, 1990) (Wray, 2002; Nesselhauf, 2000) وبناء على هذا المفهوم ظهرت كثير من الأبحاث التطبيقية التي تحاول إثبات هذه النظرية بتطوير عدد من القوائم للمتلازمات اللفظية القائمة على الشيوع لاستعمالها كأداة تعليمية في مجال تعلم اللغات الأجنبية وكذلك المعالجة الحاسوبية للغات لتزويد ذهن متعلم اللغة الأجنبية والذاكرة الحاسوبية بهذه المعرفة، والتي يتعلمها الناطق الأصلي بطريقة طبيعية غير مباشرة في الغالب من خلال ممارسته واستعماله للغة منذ الولادة. ففي اللغة الإنجليزية على سبيل الامثال نجد عدد اً من القوائم المبنية على الشيوع والتي صارت مصدراً معتمد اً في تأليف مناهج تعليم اللغة واختبارات تحديد المستوى اللغوي وما إلى ذلك.

أما فيما يتعلق باللسانيات الحاسوبية فكثيرا من المهام الأساسية لتحليل وفهم اللغة حاسوبيا تعتبر ناقصة وغير مكتملة بدون إضافة قوائم أو معاجم حاسوبية للمتلازمات اللفظية، وتشمل هذه المهام كل مستويات التحليل اللغوي من المرحلة الصوتية والصرفية إلى مرحلة التحليل السيميائي المعنوي للغة. ومن هنا فقد لعبت هذه المعاجم والقوائم دورا أساسيا في تحسين جودة النتيجة النهائية لكثير من التطبيقات في اللسانيات الحاسوبية، على سبيل المثال (الترجمة الآلية، الاستخراج الآلي للمعرفة اللغوية، التحليل الآلي لمستويات اللغة، التعرف الآلي على الأنماط اللغوية المختلفة) وإذا ما أمعنا النظر في الأبحاث في هذا المجال نجدها مطبقة على اللغة الإنجليزية، وذلك لكثرة المصادر اللغوية عالية الجودة المتاحة للباحثين، وكذلك توفر الأدوات الحاسوبية الحديثة لمعالجة وتحليل اللغة في مختلف مستوياتها، من جهة أخرى لاتزال اللغة العربية بحاجة ملحة إلى الكثير من البحث في هذا المجال، وخاصة فيما يتعلق بالأبحاث التي تركز على استعمال المتلازمات اللفظية في اللغة المعاصرة والتي تعددت تراكيبها ودخلت فيها الكثير من الكلمات الجديدة من لغات أخرى أو من اللغة نفسها وتغير معناها الدلالي بشكل كبير، وكذلك الأبحاث التي تركز على تطوير قوائم ومعاجم حاسوبية حديثة للغة العربية مبنية على الشيوع في المدونات العربية الضخمة المطورة حديثا والتي تضم بلايين لكلمات العربية المستعملة في اللغة العربية المعاصرة.

ومن هنا جاء هذا البحث ليسهم في سد هذا العجز ويساعد على ردم هذه الجفوة في المعرفة في اللسانيات الحاسوبية العربية.

وسيقسم البحث الحالي على النحو التالي، أولاً سنحاول تقديم تعريف مبسط عملي للمتلازمات اللفظية المعنية في التجربة المقدمة في هذه الدراسة، مع إعطاء تصور مختصر عن الخلاف الموجود في الدراسات السابقة فيما يتعلق بتعريف المتلازمات اللفظية، وكذلك اختيار المصطلح الملائم لهذه الظاهرة. ثانياً: سيقدم البحث تقريرا مختصرا عن تجربة عملية لاستخراج عدد من المتلازمات اللفظية في اللغة باستخدام نموذج محوسب (n-gram_tool) العربية معتمدة على تقنية الاستخراج الآلي للمتتابعات اللفظية من كلمتين إلى ست كلمات مكون من ثلاث مراحل رئيسة، وسيختم البحث بمناقشة أهم النتائج الحالية وإعطاء تصور عن التجارب المستقبلية المحتملة المتعلقة بهذه الدراسة.

1-1 **مفهوم المتلازمات اللفظية** :

من خلال نظرة سريعة للدراسات السابقة في هذا المجال نكاد نجزم على عدم وجود إجماع بين اللسانيين والحاسوبيين على تعريف مشترك لظاهرة المتلازمات اللفظية على الرغم من إدراكهم لأهميتها ودورها المحوري في تشكيل وفهم اللغات البشرية فقد اقترحت مجموعة كثيرة من التعاريف كما في الدراسات السابقة التالية (Baldwin, 2004, Calzolari *et al*., 2002 ، Guenthner & Blanco, 2004). ومثل هذا الاختلاف يكون منطقيا إذا ما إدركنا مدى تعقيد هذه الظاهرة اللغوية وتداخلها القوي مع كل مستويات التحليل اللساني (الصوتي الصرفي والنحوي والدلالي) على سبيل المثال هناك عدد من الباحثين في اللسانيات الحاسوبية يشيرون بهذا المصطلح إلى مجموعة متشابهة من الظواهر اللغوية مثل المتلازمات الاسمية والفعلية والحرفية والحكم والأمثال المتداولة (Gralinski *et al*., 2010 ، Sag *et al*., 2002) بينما يقتصر باحثون آخرون في إطلاق هذا المصطلح على الألفاظ المتصاحبة في السياقات النصية المختلفة بناء على التحليل الإحصائي لمدونات اللغة.

فنجد في معظم الدراسات السابقة أن كل باحث حاول أن يعرف هذه الظاهرة اللسانية من الزاوية التي يوليها العناية والأهتمام في بحثه. ومن هذا الباب ولضيق المجال في الورقة الحالية عن الوصول إلى تعريف شامل وكامل لهذه الظاهرة، اعتمد الباحث على واحد من أكثر التعاريف انتشارا وأقربها دلالة على كل أنواع المتلازمات اللفظية المقصودة في البحث الحالي وهو تعريف اللسانية الانجليزية (Wray, 2002) التي عرفت في كتابها المطول عن ظاهرة المتلازمات اللفظية في اللغات البشرية بأنها الكلمات المتتابعة بشكل متصل أو مع وجود فاصل بينها والتي يظهر أنها تعالج وتحفظ في الذاكرة كجملة واحدة مسبقة التركيب خلافا للتراكيب اللغوية الأخرى التي يمكن فهمها وتحليلها بسهولة عن طريق قواعد اللغة وبواسطة فهم الكلمات المفردة المكونة لتركيب أو الجملة. فكل ما يندرج تحت هذا المفهوم من المتلازمات اللفظية هو محل عنايتنا في هذا البحث.

أما فيما يتعلق بالمصطلحات المستعملة في الدلالة على هذه الظاهرة فقد تعددت وكثرت ولا يوجد مصطلح واحد متفق عليه لوصف هذه الظاهرة ويمكن أن يفسر هذا التعدد والاختلاف بنفس التفسير السابق الذي ذكرناه بخصوص تعدد واختلاف التعريفات التي اقترحت لهذه الظاهرة، ففي اللغة الإنجليزية على سبيل المثال ذكرت (Wray, 2002) أنها عثرت على أكثر من خمسين مصطلحا كلها تشير تقريباً إلى نفس الظاهرة واقترح اللساني (Schmitt, 2010) كحل لهذه المشكلة المصطلحية اختيار مصطلح Formulaic Sequence ليكون مظلة تدخل تحتها كل أنواع المتلازمات اللفظية في اللغات وفي اللغة العربية كذلك نجد نفس الاختلاف حول المصطلح الأمثل لو هذه الظاهرة فقد أطلقت عدة مصطلحات في الدراسات السابقة مثل (المتتابعات اللفظية، المتصاحبات اللغوية، الحكم، الأمثال، الشواهد) وكخطوة عملية ضرورية اختار الباحث استعمال مصطلح لمتلازمات اللفظية هنا ليدل على كل الأنماط والتراكيب اللغوية المقصودة بالبحث في هذه الورقة.

2-1 **أهداف الدراسة ودوافعها** :

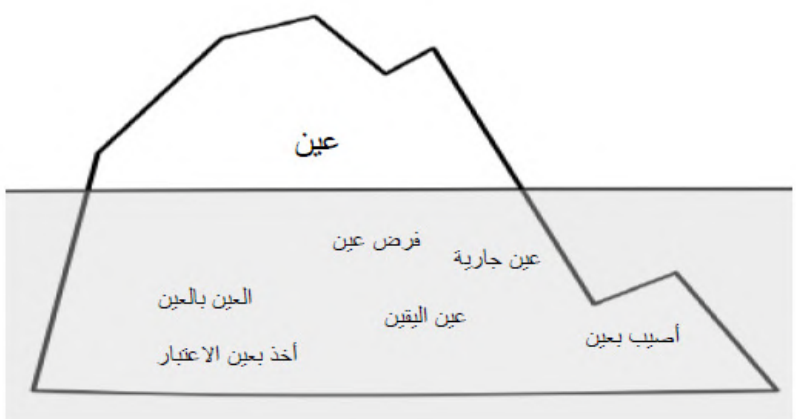
تهدف الدراسة الحالية إلى بناء معجم حاسوبي للمتلازمات اللفظية في اللغة العربية المعاصرة تحقيقا لغرضين أساسين هما:

- تطوير معجم محوسب للمتلازمات اللفظية مبني على عدد من أكبر المدونات العربية المتاحة للعربية المعاصرة يمكن الاستفادة منه في تحسين جودة المخرجات النهائية لأدوات المعالجة الطبيعية للغة العربية في كل المستويات اللسانية وتطبيقات معالجة اللغة المتنوعة(الترجمة الآلية، التعرف التلقائي للأنماط اللغوية، البحث الدلالي في النصوص، وغيرها).

- أداة تعليمية ومرجع في تعلم وتعليم اللغة العربية للناطقين بغيرها حيث يمكن الاعتماد عليها في تطوير مناهج تعليم اللغة والاختبارات المحوسبة لتحديد المستوى اللغوي بالإضافة إلى الأدوات الحاسوبية التي تساعد على تحديد مستوى صعوبة النصوص وملاءمتها للمستويات اللغوية المختلفة.

إن إهمال هذا النوع من التراكيب اللغوية في أي مهمة جادة للمعالجة الحاسوبية للغة العربية أو في مجال تعلم وتعليم اللغة العربية يؤدي إلى تدهور كبير في جودة المخرجات النهائية لهذه المهام كما دلت على ذلك كثير من الدراسات السانية والحاسوبية الحديثة (e.g., Sinclair, 1987; Martinez & Murphy, 2011) والكلمات المفردة التي يخصص لها النصيب الأكبر في الدراسات المعجمية واللسانيات الحاسوبية ماهي إلا جزء صغير من مجموعة كبيرة من التراكيب والمتلازمات اللفظية المعقدة والتي لا يمكن الوصول لفهم حقيقي للغة إلا بفهم شامل لمعانيها المتعددة في السياقات المختلفة فعلى سبيل المثال عندما يتعلم طالب اللغة العربية كلمة عين في بمعناها الحقيقي الأساسي الذي يدل على العين الباصرة يخيل له انه قد استوعب

كل معاني هذه الكلمة ويمكنه بسهولة أن يدرك معناها عندما يواجهها في سياقات أخرى لكن الحقيقة أنه سيواجه عددا من الصعوبات في فهم نفس هذه الكلمة عند تغير التركيب الذي وردت فيه لأنه إنما تعلم معنى واحد اً من المعاني متعددة لهذه الكلمة في السياقات المختلفة فالكلمة المفردة التي نتعلمها في الغالب ماهي إلا رأسا لجبل من الثلج لايظهر لنا للوهلة الأولى لكنه سرعان ما يظهر لنا عند التعمق في المحيط ونواجه الكلمة نفسها في سياقات لغوية مختلفة كما يظهر بشكل واضح في الصورة 1 التي تبين مدى تعقيد معاني كلمة عين في السياقات اللغوية المتنوعة.



صورة 1 : صورة تظهر مدى تعقيد معاني التراكيب المتعلقة بكلمة عين في اللغة العربية

2- الدراسات السابقة:

سيركز سردنا للدراسات السابقة في هذه الورقة على أهم المحاولات السابقة لبناء معاجم حاسوبية للمتلازمات اللفظية في اللغتين العربية والانجليزية وسنقوم بمقارنة مختصرة بين الدراسة الحالية وما سبقها من أبحاث في هذا المجال. فيما يتعلق بقوائم المتلازمات اللفظية التي طورت لاستعمالها أداة تعليمية في مجال تعليم اللغات الأجنبية نجد في اللغة الإنجليزية عددا من المحاولات أولها كانت محاولة اللساني (Leech *et al.* 2001) لتطوير قائمة للمتلازمات اللفظية عندما كان مشتركا في مشروع إنشاء المدونة الوطنية للغة الإنجليزية التي تتكون من مئة مليون كلمة وتعتبر أول دراسة منشورة في هذا المجال في اللغة الإنجليزية واعتمدت الدراسة على الاستخراج الآلي للعبارات المتجمدة من المدونة بعد عملية التحليل الحاسوبي الإحصائي الأولي ولذا نجد أن هذه القائمة اعتمدت على الألفاظ المتجاورة بشكل تام فقط مثل التركيب الإنجليزي (so that) والذي يمكن فهمه ككلمة واحدة مفردة وإن تكون من لفظين متتابعين وبناء على هذه النظرة الضيقة للمتلازمات اللفظية أهملت الدراسة كل أنواع المتلازمات اللفظية الأخرى والتي قد تتكون من عدة ألفاظ متقطعة وغير متتابعة أو متجمدة مثل التركيب الإنجليزي

(e.g. write down، write it down) ومثل ذا النوع من التراكيب قد يصعب استخراجه بالاعتماد على التعرف الحاسوبي الآلي البحت. وفي دراسة أخرى قام بها Durrant (2009) حاول الباحث أن يطور قائمة للعبارات الأكثر استعمالا في الكتابة الأكاديمية العلمية وكانت مبنية على مدونة للكتابات العلمية قام بتطويرها تتألف من حوالي 25 مليون كلمة وكان تركيز الدراسة في اختيار البارات على معيار الشيوع فأهملت كل العبارات اللغوية غير الشائعة بحجة عدم أهميتها وهذا قد يتسبب في فقدان مجموعة كبيرة من المتلازمات اللفظية المهمة قليلة الاستعمال. وقد اعتمد المنهج الإحصائي في النموذج الحاسوبي الذي طوره لاستخراج المتلازمات اللفظية من النصوص الأكاديمية بناء على الكلمات الأكاديمية الأكثر شيوعا في المدونة. وفي دراسة أخرى من قبل Martinez & Schmitt (2012) حاول الباحثان تطوير قائمة للمتلازمات اللفظية في اللغة الإنجليزية العامة مبنية على الشيوع ومرتبطة بشكل خاص بقائمة الكلمات الشائعة للغة الإنجليزية التي جمعها West (1953) وسماها ب general service list - GSL. وهدف الدراسة الأساسي هو استعمال هذه القائمة كأداة تعليمية في تأليف مواد لتعلم وتعليم اللغة الإنجليزية للناطقين بغيرها وقد اعتمد الباحث على الشيوع كمعيار أساسي واستعمل الأداة الحاسوبية WordSmith tools لاستخراج العبارات الأكثر شيوعا معتمدا على الدونة البريطانية الوطنية للغة الإنجليزية والتي تتكون من مئة مليون كلمة من النصوص المكتوبة والمسموعة.

أما فيما يتعلق بالدراسات العربية الحاسوبية في هذا المجال فالملاحظ بشكل عام هو قلة الدراسات بشكل ظاهر مقارنة باللغات المعاصرة الأخرى وفيما يلي سرد لأهم الدراسات التي طبقت على اللغة العربية المعاصرة في هذا المجال. من أهم الدراسات وأكثرها تداولا في هذا المجال دراسة محمد طية للمتلازمات اللفظية في اللغة العربية والتي كانت جزءا من رسالة دكتوراه تهدف إلى تعزيز وتحسين جودة المحلل الحاسوبي الصرفي الذي قام ببنائه لمعالجة المستوى الصرفي في اللغة العربية وقد وجد أن تضمين هذا النوع من العبارات اللغوية في غاية الأهمية لتحسين جودة المخرج النهائي لهذه الأداة الحاسوبية واعتمدت الدراسة في منهجيتها على طريقة شبه آلية لاستخراج المتلازمات اللفظين من مدونة للغة العربية لم تذكر أي تفاصيل عنها في الدراسة المذكورة وبناء على تصنيف سابق لهذه العبارات قام به (Sag *et al.* 2002) اعتمد محمد عطية هذه التصنيف وحاول تطبيقه على اللغة العربية فقسم هذه العبارات بحسب معانيها إلى أربعة أصناف بحسب مدى غموض معانيها ومرونة التراكيب النحوية التي تتكون منها.

وفي دراسة أخرى قام بها Hawwari *et al* (2012). حاول الباحثون بناء معجم حاسوبي مبني على عدد من المعاجم المنشورة سابقا للمتلازمات اللفظية في اللغة العربية كما في المعاجم المطبوعة التالية: (Abou Saad, 1987; Seeny *et al.*, 1996; Dawod, 2003; Fayed, 2007) وذلك لاستعماله كمادة

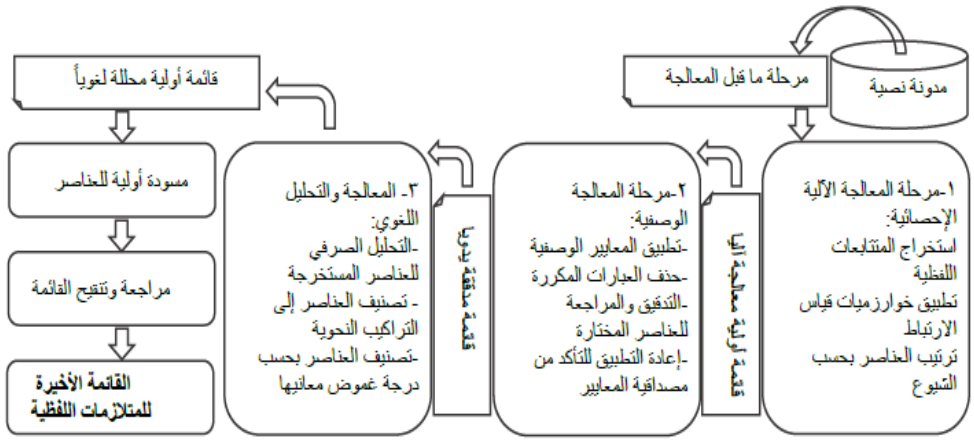
لغوية لتدريب الحاسوب على التعرف الآلي على هذه العبارات باستخدام تقنيات وخوارزميات تعلم الآلة. وفي دراسة تالية (2014) Hawwari *et al*. حاول نفس الباحثون تطوير معجم للمتلازمات اللفظية المتداولة باللهجة المصرية واقترحوا إطارا نظريا للمعالجة الحاسوبية والتصنيف اللغوي لهذه العبارات وكلا المعجمين غير منشورين حتى كتابة هذا البحث. ومن خلال هذه النظرة المختصرة للدراسات السابقة تظهر لنا أهمية دراستنا الحالية ومدى ضرورة القيام بتطوير معجم حاسوبي مبني على المدونات اللغوية المطورة حديثا للمساعدة في تحسين عدد من مهام المعالجة الحاسوبية للغة العربية وكذلك استعماله كأداة تعليمية في تطوير مواد تعلم وتعليم اللغة العربية واختبارات اللغة وغيرها من المهام التربوية اللغوية.

3- منهج الدراسة وطريقة البحث

اعتمدت التجربة الحالية على نموذج حاسوبي يجمع بين استعمال الطرق الإحصائية الآلية والوصفية اليدوية ويتكون من ثلاث مراحل أساسية يتخللها عدد من المهام الفرعية والتي تنظم عملية استخراج المتلازمات اللفظية من المدونة العربية وفي القسم التالي شرح مختصر لمنهجية استخراج المتلازمات اللفظية في الدراسة الحالية.

3-1 النموذج الحاسوبي المطبق لاستخراج المتلازمات اللفظية العربية

كما يظهر بشكل واضح في النموذج أدناه تتكون عملية الاستخراج من عدة خطوات تبدأ باختيار المدونة التي يعتمد عليها في هذه التجربة ثم تمر بمرحلة ما قبل المعالجة والتي تشتمل على تصفية نصوص المدونة من الأخطاء الإملائية وتنميط الكلمات المختلفة كتابيا مثل توحيد طريقة كتابة الهمزة وألف الوصل وحذف النصوص المكررة وما إلى ذلك بعد ذلك تبدأ مرحلة المعالجة الآلية الإحصائية والتي تتضمن الاستخراج الآلي للمتتابعات اللفظية في المدونة بناء على نموذج المتتابعات الإحصائي (**n**–gram model) وتتبع هذه المرحلة مرحلة التحليل الوصفي والتي تركز على تطبيق مجموعة من المعايير الوصفية التي تم الاعتماد عليها في اختيار التراكيب المناسبة والتي سيتم توضيحها لاحقا بالتفصيل ثم تختتم عملية استخراج المتلازمات بمرحلة التحليل اللساني الآلي واليدوي لمخرجات المراحل السابقة في هذا النموذج، وسنتناول بالتفصيل جميع مراحل هذا النموذج مع الأمثلة خلال شرح التطبيق العملي للتجربة في هذه الأقسام التالية.



نموذج 1 : نموذج لمراحل استخراج المتلازمات اللفظية من المدونة

لأن من الأهداف الأساسية للدراسة الحالية استخراج قائمة للمتلازمات اللفظية مفيدة لمتعلمي اللغة العربية لابد أن يكون معيار الشيوع من المعايير المهمة في اختيار العبارات في هذه التجربة لأن كثيرا من الدراسات في اللسانيات التطبيقية أكدت على ان معيار الشيوع من أهم المعايير التي يمكن الاعتماد عليها في تحديد مدى فائدة العبارة اللغوية لمتعلمي اللغة (e.g.، Nation, 2001; O'Keeffe *et al.*, 2007).كما أكدت كثيراً من الدراسات المطبقة على اللغة الإنجليزية أن قوائم المفردات والعبارات المبنية على الشيوع لها دور محوري في إعداد وتصميم المواد اللغوية (e.g.، Nation, 2001; O'Keeffe *et al.*, 2007; Schmitt and Martinez, 2012). وبما أن الشيوع سيكون معيارا رئيسا فلابد كذلك أن يكون عدد الكلمات في كل عبارة محدودا وقليلا لأن العبارات الشائعة غالبا ما تكون قليلة الكلمات بسيطة التراكيب وهذا ما ستركز عليه التجربة الحالية. أما فيما يتعلق بحجم القائمة أو المعجم الحالي فسيكون نطاق التجربة الحالية محدودا بحيث لا يتجاوز خمسة آلاف عبارة وذلك سبب الوقت المحدود لهذه التجربة وكذلك بناء على عدد من الدراسات السابقة التي أثبتت أن هذا الرقم مناسب لقائمة محدودة مطورة ذات أهداف محددة.

وفيما يتعلق باختيار المدونة التي سيعتمد عليها في هذه التجربة سنختار المدونة اللغوية المطورة في جامعة ليدز من قبل الباحث Sharoff (2006) و تحوي هذه المدونة أكثر من 176 مليون كلمة وتشتمل على أهم مواضيع الحياة اليومية وكذلك فيها مجموعة كبيرة من النصوص الأكاديمية العلمية وهذا الاختيار يعود لعدة أسباب من أهمها مراعاة الخطوات العلمية المتبعة أثناء جمع النصوص في هذه المدونة وكذلك تنوع المواضيع التي تندرج تحتها نصوص المدونة فنجد فيها على سبيل المثال نصوصا عن السياسة

والرياضة والفن والعلوم النظرية والتطبيقية ومن الأسباب كذلك كبر حجم هذه المدونة والتي يساعد على تمثيل مقبول إلى حد ما للغة العربية المعاصرة.

2-3 معايير اختيار المتلازمات اللفظية في التجربة الحالية:

بناء على التعريف المعتمد للمتلازمات اللفظية في هذه الدراسة وكذلك النموذج الحاسوبي المقترح سيكون اختيار العبارات في المرحلة الثانية من هذه التجربة مبني على عدد من المعايير الوصفية التالية:

أ– أن يكون في العبارة نوعا من الغموض الدلالي بحيث أنه لا يمكن بوضوح فهم معنى العبارة من خلال هم معاني كلماتها فمثلا في عبارة انتقل إلى رحمة الله قد يوجد بعض الإشكال في فهمها من خلال فهم محدود للكلمات المكونة لها ويمكن معرفة هذا النوع من العبارات بطريقة استبدال العبارة بكلمة واحدة تدل على معناها كما في العبارة السابقة يمكننا استبدالها بالفعل مات أو توفي.

ب– أن لا يكون للعبارة معنى في حد ذاتها إلا إذا وجدت في سياق معين وأغلب هذه العبارات في اللغة العربية تتكون من الأدوات

كحروف الجر والنصب كما في عبارة على الرغم من وكذلك عبارة من أجل أن وغيرها من التراكيب المشابهة.

ت– أن يكثر استعمال العبارة في الدلالة على وظيفة معنوية أو اتصالية مثل العبارات المتكررة المشهورة في اللغة كألفاظ التحيات والشكر والمجاملة (السلام عليكم، مع السلامة، شكرا جزيلاً...)

4- تطبيق التجربة:

الهدف من التجربة الحالية هو التأكد من مدى فاعلية النموذج الحاسوبي المقترح لاستخراج المتلازمات اللفظية وكذلك تعتبر خطوة أولية لجمع مادة لغوية لبناء معجم حاسوبي شامل للمتلازمات اللفظية في العربية المعاصرة مبني على المدونات اللغوية. كما تم ذكره سابقا في قسم المنهجية في هذا البحث يعتمد تطبيق هذا النموذج الحاسوبي على عدد من المهام والمراحل التي طبقت بالتفصيل كما يلي:

1-4 الخطوات العملية وإجراءات تطبيق التجربة:

1-1-4 اختيار المدونة ومهام ما قبل المعالجة:

في هذه المرحلة تم الاعتماد على واحدة من أهم مدونات اللغة العربية المعاصرة والتي طورت في جامعة ليدز وبدأت بعد ذلك مهام ما قبل المعالجة التي اشتملت على المعالجة الحاسوبية لتنقية وتصفية المدونة من الأخطاء الكتابية وتفادي التكرار بتوحيد طريقة كتابة بعض الحروف العربية كحرف الألف الذي له عدد من الأشكال وذلك لتهيئة المدونة للمرحلة التالية.

2-1-4 مرحلة المعالجة الإحصائية الآلية :

من خلال استعمال أداة الاستخراج الآلي للمتتابعات اللفظية (n-grm tool) وبنافذة محددة من كلمتين إلى أربع كلمات لكل تركيب تم بطريقة آلية استخراج أكثر من خمسة آلاف عبارة في هذه المرحلة بحيث لا يقل شيوع كل تركيب عن عشر مرات في كل مليون كلمة واختيار هذا الرقم مبني على ما وجده الباحث من توصيات سابقة في الدراسات السابقة (Church & Hanks, 1990; Stubbs, 1995). الجدول التالي يوضح عدد من العبرات المستخرجة والمعلومات الإحصائية المرتبطة بها مرتبة بحسب درجة شيوعها في المدونة.

متلازمات لفظية ثنائية	الشيوع	خوارزميات الارتباط			
		T-score	MI	Log likelihood	Log Dice
بشكل	68964	258.296	5.182	451,430	9.208
من خلال	59900	233.999	4.507	289,756	9.131
بسبب	49771	225.534	5.071	329,927	8.829
بالنسب	47091	214.019	5.407	341,640	8.667
من أجل	37889	193.911	5.378	263,642	8.566
وسائل الإعلام	5998	77.936	11.250	86,598	12.221
سبيل المثال	9452	97.299	12.252	156,386	12.924
يوم القيامة	8323	91.442	10.299	113,011	11.182

جدول 1 : أمثلة من العبارات الثنائية المستخرجة خلال المرحلة الأولى للتجربة

3-1-4 مرحلة تطبيق المعايير الوصفية :

استحوذت هذه المرحلة على الوقت الأكبر خلال إجراء هذه التجربة لأنها اعتمدت على التدقيقي اليدوي للعبارات المستخرجة من المرحلة السابقة ومحاولة تطبيق المعايير الوفية المحددة مسبقا وذلك لتنقية القائمة المطورة من كل التراكيب غير المرغوب فيها وتحسين جودة العناصر المستخرجة لتوافق الأهداف المحددة لهذه التجربة. وقد حذف خلال هذه التجربة عدد كبير من العبارات المستخرجة وذلك لمخالفتها للمعايير المعتمد عليها لاختيار المتلازمات اللفظية في هذه الدراسة ومن الأمثلة على العبارات المستبعدة ما يلي:

– العبارات التي تدل على اختصارات أو كلمات غير مفهومة أو أسماء أعلام وشركات أو أرقام مكتوبة.

– العبارات المكررة والتي تستعمل للدلالة على نفس المعنى مع اختلاف بسيط في طريقة التركيب.

– العبارات التي لا يصح أن تمثل اللغة العربية المعاصرة وذلك لاحتوائها على ألفاظ غير عربية أو محاكاة لأحد اللهجات العربية المحلية.

– العبارات المكونة من عدد من الحروف وليس لها معنى مستقل واضح كتركيب مفيد مثل من إلى وغيرها.

– العبارات التي لا يمكن أن يوجد أي غموض في فهم معناها الدلالي وذلك لتطابق معاني المفردات مع معنى التركيب الذي يتكون منها مثل عبارة وسائل الإعلام والجدول التالي يظهر بعض الأمثلة للعبرات المستبعدة في هذه المرحلة.

تعديل أي من حقول ملفك الشخصي باستثناء	اسم المستخدم	ب ف هو ثابت لك منذ تسجيلك . إذا كان الأمر
على مسرح القومي . . فقرا في بعض يعني	متن عايز	تشوف مسرحية او حتى تمدي كده من جنب المسرح
الغربة غربة الروح أكثر ما هي غربة المكان و	بالنسبة	صفات الغرب الحميدة . لا أبالغ حين أقول
ظيلة من الدولارات . و الآن يتقاضى عشرين	مليون دولار	ل يبحث عن كنز+ه القومي . ستة أسفار أشقام
واسعة في وسائل الإعلام المطبوعة . حيث ان	وسائل الإعلام	في المملكة العربية السعودية تخصص ل سيطرة
عشرة ترجمة ل نهج البلاغة . و هذا ما يوضح	إلى حد ما	مكانة الكتاب و قيمت+ه بين المسلمين . هناك
و تهدف ل زيادة طاقة التفريخ و التحميل من	وعلى	السفن . و قال بان عمليات الانشاء تمت على
. عضو هيئة التحرير ل مجلة عالم الغذاء ب	المملكة العربية السعودية	عضو لجنة الإعجاز العلمي ب المجلس الأعلى
حسناء في عيون أهل بيت+ها . . . و كل ذلك	من أجل أن	تكون وفق رضا+هم . ذ هل هي امرأة خدومة .

جدول 2 : أمثلة للعبارات المستبعدة في مرحلة تطبيق المعايير الوصفية

وفي نهاية هذه المرحلة وبعد مراجعة دقيقه لكل العبارات المستخرجة قام بها الباحث بالاعتماد على المعايير الوصفية المحددة في هذه التجربة تم استبعاد القسم الأكبر من العبارات المستخرجة في المرحلة الأولى وبقي حوالي 208 متلازمة لفظية انطبقت عليها المعايير المحددة في هذه المرحلة. وقد عمل

الباحث على تأكيد وتوثيق طريقة تطبيق المعايير الوصفية بتطبيقها مرة أخرى من قبل شخص محايد ليس له علاقة بالبحث الحالي وهو أحد الطلاب المختصين في اللسانيات وكانت النتيجة الإحصائية متقاربة في تطبيق المعايير الوصفية وهذا بدوره عزز من الاعتماد على المعايير السابقة.

كما شملت هذه المرحلة على استخراج أمثلة لكل متلازمة لفظية من المدونة وكان التركيز على الأمثلة التي تبرز الاستعمال الحي للغة العربية المعاصرة كما يظهر ي الجدول التالي:

أمثلة	المتلازمات اللفظية
كان من أبطال الفيلم الأساسيين على الرغم من بعض العثرات في تقليد اللهجة	على الرغم
التعرض ل أشعة الشمس لفترات طويلة خطر يهدد جميع الأعمار بغض النظر عن كون من يتعرض لها رضيع أو طفل أو شاب.	بغض النظر عن
أن تناول طعام صحي وسليم مع اللعب النشيط يؤدي إلى صحة جيدة. وبالتالي فأن الصحة الجيدة تؤدي إلى شهية جيدة.	وبالتالي
يجب أن تكون التكاليف في الإسلام حسب طاقة المكلف فلا يطلب منه على سبيل المثال ألا أثنين ونصف بالمائة من ربحه السنوي الصافي كزكاة.	على سبيل المثال

جدول 3 : المتلازمات اللفظية مع أمثلة استعمالها

4-1-4 مرحلة التحليل والتصنيف اللغوي :

في هذه المرحلة تم تمرير العناصر المستخرجة من المراحل السابقة على مجموعة من مراحل التحليل اللغوي أولها التحليل الصرفي الآلي والذي اشتمل على تقطيع الكلمات صرفيا ووسم كل كلمة بنوعها (اسم ـ فعل ـ حرف) ثم تحديد أجزاء الكلام لتراكيب العناصر المستخرجة وأخيرا التحليل الدلالي الذي من خلاله صنفت الكلمات إلى عدة أصناف بناء على درجة غموض المعنى في المتلازمات اللفظية. في هذه المرحلة اعتمد الباحث أداة مدى أميرة للتحليل الصرفي MADAAMIRA والجدول التالي يوضح التصنيف الصرفي الأولي للمتلازمات اللفظية المستخرجة في هذه التجربة.

أمثلة	نوع الكلمة الأولى
نظرا ل	اسم
متعلقة ب	صفة
هنا وهناك	ظرف
ما يلي	اسم موصول
يؤدي إلى	فعل
لابد	أداة
في إطار الوصول	حرف
وبالتالي	حرف عطف
سبحان الله	اسم مصدر

جدول 4 : أمثلة لأهم الأصناف التي تنتمي لها أول كلمة في المتلازمات اللفظية

أما بالنسبة لتحديد أجزاء الكلام فقد ظهر في هذه المرحلة أن المتلازمات اللفظية تنتمي إلى أغلب أنواع تركيب الجملة في اللغة العربية.

كما يظهر بالتفصيل في الجدول التالي لأهم تراكيب المتلازمات اللفظية في اللغة العربية:

أمثلة	التركيب
بمناسبة	حرف + اسم
ردا على	اسم + حرف
في نهاية المطاف	حرف+ اسم + اسم
سبحان الله	اسم مصدر + اسم
وبالتالي	حرف عطف + حرف جر + اسم
رحمه الله	فعل + ضمير + اسم
من الضروري	حرف + صفة
مرتبط ب	صفة + حرف
هنا وهناك	ظرف + حرف عطف + ظرف

جدول 5 : أمثلة لبعض التراكيب النحوية التي صنفت لها المتلازمات اللفظية في التجربة

وركزت الخطوة الأخيرة في التحليل اللغوي على محاولة تصنيف المتلازمات اللفظية إلى عدة فئات بناء على درجة غموض المعاني التي تدل عليها فكانت هناك ثلاث فئات في هذا المجال غموض شديد ومتوسط وبسيط ومع وجود تداخل قوي في هذا التصنيف إلا نه مفيد جدا وخاصة عند محاولة استعمال هذه القائمة في مهام المعالجة الحاسوبية للغة فالباحث حينئذ لا يحتاج إلا للعبارات الأكثر غموضا والتي لا يمكن بأي حال من الأحوال فهم معانيها من خلال فهم معاني الكلمات المكونة لها.

5 - ملخص وخاتمة

في هذه التجربة المختصرة حاولنا تطبيق نموذج حاسوبي معتمد على الطرق الإحصائية والوصفية لاستخراج قائمة للمتلازمات اللفظية في اللغة العربية المعاصرة وذلك كخطوة أولية لبناء معجم حاسوبي شامل لهذا النوع من العبارات بحيث يمكن الإفادة منه في مجال معالجة اللغة العربية حاسوبيا وكذلك تصميم وإعداد المواد التعليمية للغة العربية وخاصة لغير الناطقين بها. وقد اشتمل النموذج المطبق على ثلاث مراحل رئيسة يتخللها مجموعة من المهام والتي أسهمت أخيرا في إنشاء قائمة مراجعة ومنقحة للمتلازمات اللفظية.

وفي التجارب المستقبلية نرجو أن نتمكن من توسيع نطاق التجربة الحالية لتطبق على المدونات العربية الضخمة وكذلك محاولة تطبيق التجربة بطريقة آلية في كل الخطوات وذلك لتوفير الوقت والجهد وزيادة الدقة في مراحل استخراج المتلازمات اللفظية. وكذلك يمكن أن يكون هذا المعجم نواة لبيئة حاسوبية متكاملة لتعليم العربية على شبكة المعلومات الإنترنت.

المراجع :

احمد ابو سعد، & ظبية عبد الله محمد السليطي (1991). معجم التراكيب والعبارات الاصطلاحية العربية القديم منها والمولد، دار العلم للملايين.

Ackermann, K. and Chen, Y.-H. (2013). Developing the Academic Collocation List (ACL) – A corpus-driven and expert-judged approach. Journal of English for Academic Purposes. 12(4), pp.235-247.

Attia, M. (2006). Accommodating Multiword Expressions in an Arabic LFG Grammar. In T. Salakoski et al. (Eds.): Advances in Natural Language Processing. FinTAL 2006, Lecture Notes in Computer Science. Vol. 4139, pp. 87 - 98, 2006. Springer-Verlag Berlin Heidelberg.

Attia, M., Tounsi, L., Pecina, P., van Genabith, J., & Toral, A. (2010). Automatic extraction of Arabic multiword expressions.

Baldwin, T. (2004). Multiword Expressions, an Advanced Course. Paper presented at The Australasian Language Technology Summer School (ALTSS 2004), Sydney, Australia.

Biber, D., Conrad, S., & Cortes, V. (2004). If you look at…: Lexical bundles in university teaching and textbooks. Applied linguistics, 25(3), 371-405.

Biber, D., Johansson, S., Leech, G., Conrad, S., Finegan, E., & Quirk, R. (1999). Longman grammar of spoken and written English (Vol. 2). MIT Press.

Boers, F., Eyckmans, J., Kappel, J., Stengers, H., & Demecheleer, M. (2006). Formulaic sequences and perceived oral proficiency: Putting a lexical approach to the test. Language teaching research, 10(3), 245-261.

Butt, M. (1999). A Grammar Writer's Cookbook. Stanford, CA: CSLI.

Calzolari, N., Lenci, A., & Quochi, V. (2002). Towards Multiword and Multilingual Lexicons Between Theory and Practice. na.

Capel, A. (201)0. A1–B2 vocabulary: insights and issues arising from the English Profile Wordlists project. English Profile Journal. 1(01), pe3.

Church, K.W. and Hanks, P. (1990). Word association norms, mutual information, and lexicography. Computational linguistics. 16(1), pp.22-29.

Coxhead, A.(2000). A new academic wordlist. TESOL Quarterly, 34(2), pp.213-238.

Davies, M. and Gardner, D. (2013). A Frequency Dictionary of American English: Word Sketches, Collocates and Thematic Lists. Routledge.

Dawood, M. (2003). A Dictionary of Arabic Contemporary Idioms (mu'jm alta'bir alastlahiat). Dar Ghareeb.

Diab, M.T. and Krishna, M. (2009)a. Handling sparsity for verb noun MWE token classification. In: Proceedings of the Workshop on Geometrical Models of Natural Language Semantics: Association for Computational Linguistics, pp.96-103.

Diab, M.T. and Krishna, M. (2009)b. Unsupervised classification of verb noun multi-word expression tokens. Computational Linguistics and Intelligent Text Processing. Springer, pp.98-110.

Dipper, Stefanie. (2003). Implementing and Documenting Large-Scale Grammars – German LFG, Institut f¨ur maschinelle Sprachverarbeitung, Institut f¨ur maschinell Sprachverarbeitung der Stuttgart University: Ph.D.

Dorgeloh, H. and Wanner, A. (2009). Formulaic argumentation in scientific discourse. Formulaic language. 2, pp.523-44.

Durrant, P. (2009). Investigating the viability of a collocation list for students of English for academic purposes. English for Specific Purposes. 28(3), pp.157-169.

Durrant, P.L. (2008). High frequency collocations and second language learning. thesis, University of Nottingham.

Ellis, N. C., Simpson Vlach, R. I. T. A., Maynard, C. (2008). Formulaic language in native and second language speakers: Psycholinguistics, corpus linguistics, and TESOL. Tesol Quarterly, 42(3), 375-396.

Ellis, N.C. (1996). Working memory in the acquisition of vocabulary and syntax: Putting language in good order. The Quarterly Journal of Experimental Psychology: Section A. 49(1), pp.234-250.

Ellis, R.S.-V.a.N.C. (2010). An Academic Formulas List: New Methods in Phraseology Research Applied Linguistics. 31, pp.487-512.

Erman, B. and Warren, B. (2000). The idiom principle and the open choice principle. text-the hague then amsterdam then Berlin-. 20(1), pp.29-62.

Fayed, Wafaa Kamel. (2007). A Dictionary of Arabic Contemporary Idioms (mu'jm alta'bir alastlahiat).Abu Elhoul.

Fellbaum, C. (1998). WordNet. Blackwell Publishing Ltd.

Firth, J. (1951). Papers in Linguistic [s J]. Oxford University Press.

Garside, Roger. (1987). The CLAWS word-tagging system. In R. Garside, G. Leech and G. Sampson (Eds.), The Computational Analysis of English (30-41). London: Longman.

Gralinski, F., Savary, A., Czerepowicka, M., & Makowiecki, F. (2010). Computational Lexicography of Multi-Word Units: How Efficient Can It Be?. In Workshop Multiword Expressions: from Theory to Applications.

Guenthner, F., & Blanco, X. (2004). Multi-lexemic expressions: an overview. Lexique, Syntaxe et Lexique-Grammaire. Papers in Honour of Maurice Gross, 239-252.

Habash, N., & Rambow, O. (2005). Arabic tokenization, morphological analysis, and part-of-speech tagging in one fell swoop. In Proceedings of the Conference of American Association for Computational Linguistics (pp. 578-580).

Hawwari, A., Attia, M., & Diab, M. (2014). A framework for the classification and annotation of multiword expressions in dialectal arabic. ANLP 2014, 48.

Hawwari, A., Bar, K., & Diab, M. (2012). Building an Arabic multiword expressions repository. Proc. of the 50th ACL, 24-29.

Hunston, S. (2002). Corpora in applied linguistics. Cambridge University Press.

Hyland, K.(2008). As can be seen: Lexical bundles and disciplinary variation. English for specific purposes. 27(1), pp.4-21.

Kjellmer, Goran. (1990). A mint of phrases. In K. Aijmer & B. Altenberg (Eds.), English corpus linguistics: Studies in honour of Jan Svartvik (pp. 111-127). London: Longman.

Lee, D.(2002). Notes to accompany the BNC WORLD edition (bibliographical) index. Unpublished manuscript, last edited. 30.

Leech, G., & Rayson, P. (2014). Word frequencies in written and spoken English: Based on the British National Corpus. Routledge.

Leech, G., Garside, R., & Atwell, E. S. (1983). The automatic grammatical tagging of the LOB corpus. ICAME Journal: International Computer Archive of Modern and Medieval English Journal, 7, 13-33.

Li, W., Zhang, X., Niu, C., Jiang, Y., & Srihari, R. (2003, July). An expert lexicon approach to identifying English phrasal verbs. In Proceedings of the 41st Annual Meeting on Association for Computational Linguistics-Volume 1 (pp. 513-520). Association for Computational Linguistics.

Liu, D. (2012). The most frequently-used multi-word constructions in academic written English: A multi-corpus study. English for Specific Purposes. 31(1), pp.25-35.

Martinez, R. (2011). The development of a corpus-informed list of formulaic sequences for language pedagogy. thesis, University of Nottingham.

Martinez, R. and Murphy, V.A. (2011). Effect of frequency and idiomaticity on second language reading comprehension. TESOL Quarterly. 45(2), pp.267-290.

Martinez, R. and Schmitt, N. (2012). A Phrasal Expressions List. APPLIED LINGUISTICS. 33(3), pp.299-320.

Meghawry, S. (2015). Semantic extraction of Arabic multiword expressions.

Mel'ćuk, I.(1998). Collocations and lexical functions. 2001 [1998]. pp.23-54.

Milton, J. (2009). Measuring second language vocabulary acquisition. Multilingual Matters.

Nation, I.S.P. (2001). Learning vocabulary in another language. Cambridge: Cambridge University Press.

Nation, P. and Waring, R. (1997). Vocabulary size, text coverage and word lists. Vocabulary: Description, acquisition and pedagogy. 14, pp.6-19.

Nerima, L., Seretan, V., & Wehrli, E. (2003, April). Creating a multilingual collocation dictionary from large text corpora. In Proceedings of the tenth conference on European chapter of the Association for Computational Linguistics-Volume 2 (pp. 131-134). Association for Computational Linguistics.

Nesselhauf, N. (2005). Collocations in a learner corpus. Amsterdam: John Benjamins.

Ohlrogge, A. (2009). Formulaic expressions in intermediate EFL writing assessment. Formulaic language. 2, pp.387-404.

O'keeffe, A., McCarthy, M., & Carter, R. (2007). From corpus to classroom: Language use and language teaching. Cambridge University Press.

Pawley, A., & Syder, F. H. (1983). Two puzzles for linguistic theory: Nativelike selection and nativelike fluency. In J. C. Richards & R. W. Schmidt (Eds.), Language and communication (pp. 191-226). New York: Longman.

Roth, R., Rambow, O., Habash, N., Diab, M., & Rudin, C. (2008, June). Arabic morphological tagging, diacritization, and lemmatization using lexeme models and feature ranking. In Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies: Short Papers (pp. 117-120). Association for Computational Linguistics.

Sag, I. A., Baldwin, T., Bond, F., Copestake, A., & Flickinger, D. (2002, February). Multiword expressions: A pain in the neck for NLP. In International Conference on Intelligent Text Processing and Computational Linguistics (pp. 1-15). Springer Berlin Heidelberg.

Sawalha, M. and Atwell, E. (2011). Accelerating the Processing of Large Corpora: Using Grid Computing Technologies for Lemmatizing 176 Million Words Arabic Internet Corpus. Advanced Research Computing Open Event.

Schmitt, N. (2004). Formulaic sequences: Acquisition, processing, and use. John Benjamins Publishing.

Schmitt, N. and Martinez, R.(2012). A Phrasal Expressions List. Applied Linguistics. 33(3), p299.

Schmitt, N.(2010). Researching vocabulary: a vocabulary research manual. Basingstoke: Palgrave Macmillan.

Scott, Mike. (1999). WordSmith Tools users help file. Oxford: Oxford University Press.

Sharoff, S. (2006). Creating general-purpose corpora using automated search engine queries. WaCky. pp.63-98.

Shin, D.a.N., P. (2008). Beyond single words: The most frequent collocations in spoken English. ELT Journal 62(4), pp.339-348.

Sieny, Mahmoud Esmail, Mokhtar A. Hussein and Sayyed A. Al-Doush. (1996). A contextual Dictionary of Idioms (almu'jm alsyaqi lelta'birat alastlahiah). Librairie du Liban Publishers.

Sinclair, J. (1987). Looking up: an account of the COBUILD Project in lexical computing and the development of the Collins COBUILD English language dictionary. London: Collins ELT.

Sinclair, John M. (1987). Collocation: a progress report. In R. Steele & T. Threadgold (Eds.), Language topics: Essays in honour of Michael Halliday (Vol. 2, pp. 319-331). Amsterdam: John Benjamins.

Siyanova-Chanturia, A. et al.(2011). Adding more fuel to the fire: An eye-tracking study of idiom processing by native and non-native speakers. Second Language Research. 27(2), pp.251-272.

Smadja, F. et al. (1996). Translating collocations for bilingual lexicons: A statistical approach. Computational linguistics. 22(1), pp.1-38.

Stubbs, M. (1995). Collocations and semantic profiles: on the cause of the trouble with quantitative studies. Functions of language. 2(1), pp.23-55.

Wray, A. (2002). Formulaic language and the lexicon. Cambridge University Press Cambridge.

Wray, A. (2004). 'Here's one I prepared earlier': Formulaic language learning on television. In N. Schmitt (Ed.), Formulaic sequences: acquisition, processing and use (pp. 249-268). Amsterdam: John Benjamins.

Wray, A. (2008). Formulaic language: Pushing the boundaries. Oxford: Oxford University Press.

Wray, A. (2009). Identifying formulaic language: Persistent challenges and new opportunities. Formulaic language. 1, pp.27-51.

Wray, A. (2013). Formulaic language. Language Teaching. 46(03), pp.316-334.

Wray, A. and Namba, K. (2003). Formulaic language in a Japanese-English bilingual child: a practical approach to data analysis. Japan Journal for Multilingualism and Multiculturalism. 1(9), pp.24–51.

TICAM'16, p. 197 - 207

Tamazight E-Learning Platform: Leveraging Graph Databases for Flexible Design

Violetta Cavalli-Sforza Arab Issar Mohamed Ouakrime
Abderrahim Agnaou Hassan Darhmaoui
Al Akhawayn University
{v.cavallisforza,a.issar,mo.ouakrime, a.agnaou, h.darhmaoui}@aui.ma

Abstract

We describe the progress made to date on the development of an e-learning platform for the Amazigh language in the context of a collaboration between Al Akhawayn University in Ifrane and the Institut Royal de la Culture Amazighe in Morocco. We focus on a novel implementation of various aspects of the course contents using graph databases. We motivate our choice of technology by discussing how current information and communication technologies could be further exploited to differentiate e-learning experiences from conventional face-to-faceinstruction.

1. Introduction and Motivation

A Tamazight e-learning platform is currently being developed at Al Akhawayn University in Ifrane (AUI) in the framework of a collaborative effort between AUI and the Institut Royal de la Culture Amazighe (IRCAM).Theover-arching goal of the project is to offer to the general adult public a carefully crafted resource for learning the Amazigh language[1].

A brief search of the web shows that there is no lack of online resources for learning the Amazigh language, some developed by generous amateurs who wish to share their language with the web community, a few more professionally crafted[2]. They use a combination of media, frequently videos, and other e-learning tools.They target specific regional variants, but often content appears to be ad hoc and not basedon sound language pedagogy.

Our goal, in contrast, is to create a platform that will serve the needs of different types of learners using a framework that lends itself to flexible expansion. The content and philosophy

1 In this paper we use the word 'Tamazight' to both refer to the Amazigh language as a collection of variants, and the variant of the language spoken in the Middle Atlas, the central region of Morocco. The context of use should make it clear whether we are referring to the Amazigh languages collectively or the Middle Atlas variant specifically.

2 The following are just a few of the available choices online for adults:
http://mylanguages.org/learn_tamazight.php,
https://www.mylanguageexchange.com/Learn/Berber.asp,
http://store.instantimmersion.com/tamazight/level-1-tamazight-online-class.html
http://www.morocco.com/forums/berber-culture-culture-berbere/24936-learn-amazigh-language-online-free.html, https://www.youtube.com/watch?v=oWuBaThHr9Y,
Among the more professionally crafted ones, we might include https://www.livelingua.com/project/peace-corps/Tamazight/

of our platformoriginatesin the extensive experience of two institutions in in language learning pedagogy. The implementation aims to make use of state-of-the-art information and communication technologies. More specifically, the project aims to: 1) provide a user experience that goes beyond the typical online course offered throughcourse management systems (CMS),learning management systems (LMS) or other online content delivery systems in order to support different types of learners and different purposes for learning; and 2) exploit graph databases to represent course contents at different levels, leveraging the flexible structure of this type of database in support of the first sub-goal. We elaborate on these pointsbelow.

The paper is organized as follows. We begin by describing the overall architecture of the system, delving in some detail on the choice of tools and the motivation for that choice. We then discuss the content of thecourse and the resources on which it is based, and follow that with a description of how the course content is being implemented with the chosen tools. We conclude with an overview of the status of the system and some remarks about our plans for the interface to the platform.

2. System Architecture and Technology Choices

We begin by examining current online course practices and technologies and why we consider them limiting. We then consider how conventional and newer technologies can be combined to offer a different user experience that is more appropriate for different language learning styles and needs.

2.1. Beyond Typical Online ContentDisplay

These days there is no shortage of online learning opportunities, including fully open and free courses (for example MIT OpenCourseWare), paying degree programs offered by educational institutions in parallel with traditional classroom instruction, and sites such as Coursera and Udacity, some of whose courses can be taken at no charge but require payment if a certificate or diploma is desired. These courses are offered through course management systems (CMS), some open source like **moodle**[3] and **OpenedX**[4], some proprietary like Jenzabar[5] or Blackboard[6], which differ in look-and-feel but fundamentally offer the same typesof functionality: the ability to store, post and display different types of content (text and other media); to create and post different types of assignments and tests (for instructors); to take tests and submit assignments (for students); to store information about student's progress through a course and grades.

3 https://moodle.org/
4 https://open.edx.org/
5 https://www.jenzabar.com/
6 http://anz.blackboard.com/sites/international/globalmaster/

Material is typically organized by topic and chronologically, much as you might expect to find in a textbook or course pack. The main difference is that the material is available in digital rather than paper form and, for that reason, is inherently more easily changed and updated. When it comes to tests, particularly multiple-choice or true/false questions, the systems might draw from a large database of questions allowing different students to receive different but comparable tests and, if multiple attempts at a test are permitted for the same student, different but comparable tests at each attempt.

The reader familiar with online courses should be able to recognize that, by and large, most online courses do little more than put online, in digital form the conventional course. The course may include a set of documents that the student should read or, more commonly, a series of videos that take the place of a live lecture. Direct interaction with the instructor is replaced by forums and other synchronous or asynchronous chatting mechanisms, which allow students to communicate among themselves and with teaching assistants, if not with the instructor. Depending on the type of content the course delivers, testing may be in the form of multiple choice questions, which are immediately corrected by the system, or assignments that the student has to carry out and submit, to be graded offline by the teaching team for the course or, in some cases, by the other students enrolled in the course.

The above remarks are not intended to be critical of current online courses; indeed the wealth of accessible online content is a boon for anyone who wants to keep learning but needs to do so with a minimum of temporal and geographical constraints. And LMS systems provide a number of useful tools for course developers, instructors and students. The remarks do, however, suggest that perhaps we are not exploiting our digital online content delivery systems as fully as we could in order to give more flexible access to the material, supporting learners with different styles of and different motivation for learning.Moreover, knowledge, including language knowledge, is linked in multiple ways in the learners' head: shouldn't our digital instructional platforms in some way reflect that multiple linking?

In the Tamazight E-Learning platform, we plantocater to the traditional language learner who wants to be led through a curriculum by a teacher. But, in time, we also aim to enablethe linguistically curious learners who want to explore the rich regional variation of the Amazigh language,navigating the content offered in a less linear fashion and taking advantage of the multiple types of links between materials that digital media offer. Finally, we also envision that the platform should be able to support the language learner who has a very specific and focused purpose, such as quickly developing vocabulary and communicative ability in a particular regional variant and topic area, for example the medical or legal domain.

2.2. GraphDatabases for Content Organization

Rather than developing a special course from scratch for each type of user, we can organize the content of the course from the viewpoint that each element is a resource that can be picked up and used in different places and ways as needed. From this perspective, choosing tools that allow flexible access to resources becomes acrucial aspect of platform design. The use of a graph database (GDB)is a focal feature of the Tamazight e-learning platform because of the flexibility it provides in changing the structure of information stored in the database [1].

In the widely used conventional relational databases (RDBs), links between data are stored in the data themselves through primarykeys and foreign keys. The data stored in RDBs is expected to have a good fit to a tabular format, with a fixed set of columns in a table and relationships between tables encoded in one of the columns via keys. In GDBsdata is represented by nodes that have properties (or tags) and are directly related to other nodes by edges representing semantic relationships between nodes. GDBs by design allow flexible addition of relationships between data and fast retrieval of complex structures that are difficult to model in RDBs. The actual implementation of GDBs varies: some are based on a relational implementation, but most are NoSQL based. Our choice is Neo4j, which is available under GPL v.3 and is the most popular GDB as of March 2016. It is written in Java and is accessible from various languages including Java, .NET, JavaScript and Python via APIs, in addition to other useful tools.

The GDB is used as the framework to represent the structure of the curriculum at the upper levels, as well as the structure of individual curriculum units, the dictionary, and other resources. A real advantage offered by a GDB is the ability to have a varied link structure that can differ from node to node (nodes representing similar types of objects don't have to have all the same links). The link structure does not have to be strictly designed beforehand anddoes not have to undergo a complex update process if the structure of information represented needs to change.In addition, because the GDB represents the structure of learning material, and because that node and edge structure can be queried, it can also be used to inform the Graphical User Interface (GUI) of the information to display at each level.

While it is true that GDBs may have high access times when queries require traversal of several links over a large quantity of data, heavy querying of this sort is not a common or significant operation for the e-learning platform application.

2.3. Combining MOOC Platforms, Graph DBs and Web Interfaces

Earlier, we expressed our dissatisfaction with standard platforms and LMS systems that support MOOCs. That attitude was based on the observation that they do little more than present instructional content digitally without fully taking advantage of the flexibility and opportunities for multiple linking of resources offered by a dynamic platform such as the web, as opposed to paper or other fixed media. Here we aim to indicate how existing tools can be combined to provide a richer experience for interacting with course content that is more suitable for language learners.The overall architecture of the system, as envisioned, is shown in Figure 1. The links are annotated with the information transmitted between system components from the perspective of the user.

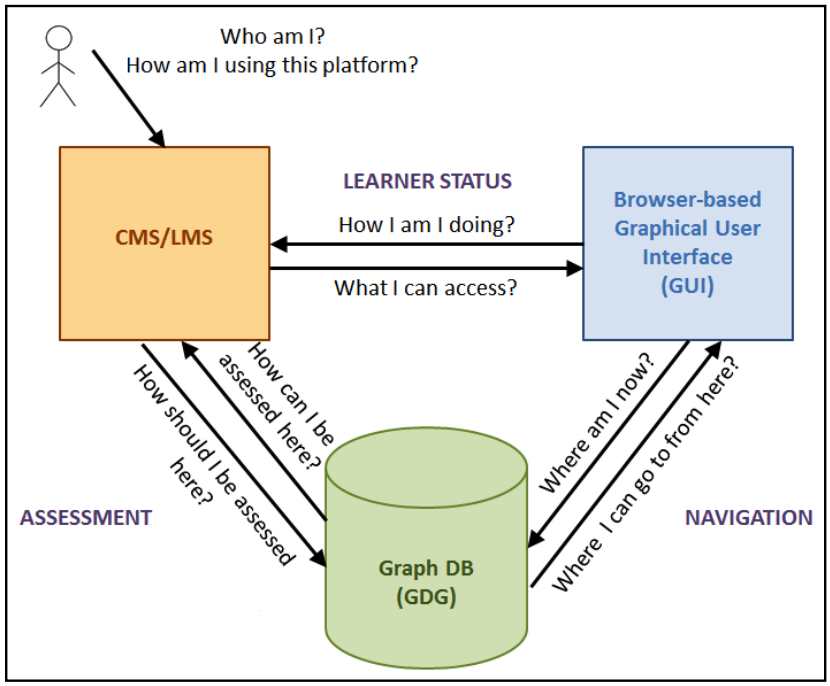


Figure 1. System Architecture

As can be seen in the architecture diagram, the Tamazight e-learning platform does not do away with using a traditional LMS or CMS; AUI uses Open edX for its small but growing e-learning offerings[7]. An LMS offers a number of useful components, among which storage of student-specific information: login, performance on different types of assessments, progress through the course, etc. We neither want nor need to re-implement these types of tools. Rather, we focus our efforts on representing the course content using the Neo4j graph DB and interacting with the DB through a dynamic web interface to navigate the graph edges.

As shown in Figure 1, the interaction between the browser-based GUI and the GDBdetermines where the learner is situated in the curriculum graph and which nodes are immediately accessible, barring access restrictions (if the course uses them). By providing a map of the curriculum and where the user is located in it, the graph can also be used to help the user navigate back to previous nodes without getting lost in complex linkages, much as a road map is used to navigate unknown areas.

The connection between the user interface and the CMS/LMS is primarily used to provide to the UI information about access restrictions for a particular learner if they are enrolled in a course that limits access to some information. That freedom to navigate may be curtailed in special circumstances. These include: blocking access to a part of the course content because it is still under construction; limiting access to content that the instructor considers too advanced

7 http://mooc.aui.ma/

for students; or controlling access to evaluative assessments that should be delivered at specific times. If the Tamazight course is delivered as a credit-bearing or certificate-awarding course, which requires instructor supervision and evaluation, student progress and performance will be managed through Open edX facilities, as will the results of formative and evaluative assessments. In contrast, if (prospective) users just wish to explore the content of the course, they may be able to do so independently of the LMS, or be asked to register through the LMS just for the purposes of tracking access, but once "in", access to different parts of the content need not be managed through the LMS. Starting from the entry point of the curriculum graph, or from any other remembered bookmarked node in the graph, the user should be able to freely navigate to any other point and back using the structure of the graph.

Finally, the connection between the GDB and the CMS/LMS serves to inform the latterabout the overall structure (not the details) of the curriculum. The CMS/LMS is used to store assessments for those aspects of the curriculum that do have associated assessments, and to provide to the GDB assessments appropriate to each component of the curriculum. If a student undergoes an assessment more than once for the same component of the curriculum, the CMS/LMS holds the information needed to determine which components of the assessment should or should not be repeated. In other words, the assessment nodes in the GDB are really placeholders for assessment content that is chosen by in consultation with the CMS/LMS based on previous history of assessments for the specific learner.

3. Planned Course Contents

3.1. Curriculum and Resources

The curriculum offered by the e-learning platform is based on a textbook of functional Tamazight – the Middle Atlas variety – that was developed by Professors Mohamed Ouakrime and the late Bouazza Bagui for a course that was offered at Al Akhawayn University (AUI) [2]. Compilation of this book, henceforth referred to as the "AUI textbook", began in the early 1990s and used the Arabic script in its first edition. The course was taught for some years at AUI and then discontinued for lack of demand. A renewal of interest in the Amazigh language on the part of students brought the course back in the fall semester of 2011. Since then, it has been offered regularly every fall and spring semester to approximately 20 students per offering. Most are Moroccan, many with Amazigh family background but others simply interested in learning the language made official by the 2011 Constitution of Morocco [3]. The latest edition of the textbook, which uses the Tifinagh script, is from 2011, but is currently undergoing revisions in response to feedback from IRCAM language specialists. More details concerning the textbook and its use in the e-learning platform will be presented below.

A second existing resource we are drawing from is the textbook for adult learners developed by IRCAM, and specifically the Central Morocco version of the volume "Sawlat s Tmazight" [4] ("Sawlat" for short). This textbook also needs to undergo revision to correct errors known to exist in its first edition, but we are planning on using some of the dialogues, activities, and other elements to enrich the limited content of the AUI textbook. It should be remembered that the AUI textbook was intended to be used by instructors in a courseand therefore provides the

backbone for a lessons but needs to be enriched with classroom activities. Sawlat is richer at the level of content, activities, and assessments, but is fully written in Tifinagh and contains very little in the way of translation, so access to its content must be mediated by an instructor. Finally, while the AUI textbook does have some recordings available to accompany the textbook, Sawlat has none, to our knowledge. We are also planning to develop more materials to provide further practice, some variation and more opportunities for learner self-expression, while remaining closely linked to the materials already available through the two textbooks.

3.2. New Resources under Construction

While the AUI and Sawlat textbooks provide a solid basis for the curriculum, there are aspects of language learning that are only barely touched or not present at all in the two books and that we are in the process of developing.

Writing and pronouncing Tamazight sounds. The first and most fundamental of the planned additions is an introductory unit that will focus on writing and pronunciation. The AUI textbook relies on in-class presentation of the alphabet and sounds of Tamazight, with a little individual in class pronunciation and at-home writing practice prior to embarking on Unit 1, which already presents simple dialogues. Sawlat contains a table with the alphabet, the names of the letters, the corresponding sounds in Latin and Arabic alphabets(when there is one) and an example of the letter's use in a word (but no translation is provided). In other words, there is a need to develop materials that provide practice on writing letters and pronouncing sounds, together with very basic reading and listening skills.

The approach we are taking to developing this unit is two-pronged. Concerning the learning of individual letters and sounds, we divide the sounds into the following categories, reflecting a bias towards learners who are not native speakers of Arabic and for which certain sounds are distinctly difficult to pronounce[8]:

1) vowels and semi-vowels;
2) "easy" consonants, that is"easy" from the perspective of AUI's foreign students, frequently from an English-speaking background, to which AUIs course is also partially addressed: ⴱ 'b', ⴼ 'f', ⵊ 'j', ⵍ 'l', ⵎ 'm', ⵏ 'n', ⵛ 'sh';
3) "paired" consonants, which mostly include pairs of consonants with non-emphatic and emphatic members such as (ⴷ'd', ⴹ 'ḍ'), (ⵜ't', ⵟ'ṭ'), (ⵔ'r', ⵕ'ṛ'), (ⵙ's', ⵚ'ṣ')and (ⵣ'z', ⵥ'ẓ'), but also (ⴳ 'g', ⴳⵯ 'gw') and (ⴽ 'k', ⴽⵯ 'kw').
4) "difficult" consonants, includingsix sounds that are present in Arabic but generally not in the native languages of AUI's most frequent visitors: ⵀ 'h', ⵃ 'ح', ⵅ 'خ', ⵄ 'ع', ⵖ 'غ', ⵇ 'ق'.

In addition to choosing to present letters and sounds in a particular sequence, we are also drawing from the phonics approach to reading [5]. This approach is extensively used for teaching reading to children and illiterate adults, entails focusing early in the curriculum on

8 Note that, thinking of course elements as a set of resources, the particular choice of sequence for presenting letters would not be hard to change.

sight words and high frequency words of the language, in addition to phonics patterns and common word patterns of Tamazight.

Where there is a choice to be made in terms of the order in which to present sounds or sound combinations, we will use frequency of occurrence in our basic corpus, the collection of text from our two textbooks.

Representing regional variation. The presence of significant regional variation in pronunciation, vocabulary, and to a lesser extent grammar is a fact of life in the Amazigh language, which has been a spoken language for several centuries. At AUI our focus is the Middle Atlas variant, specifically that of the Ifrane region and immediate surroundings (though there is variation even within the region), but we don't want to ignore the existence of variation, for two reasons. From the learner's side, there should be a recognition that the variant spoken in the region of AUI may not be exactly the same (and may be significantly different) from the variant spoken in other parts of Morocco and adjoiningamazighophone areas. From the platform developer's side, it is both wise and technically interesting to design the system to represent variation. Even if the initial focus is on the Middle Atlas variety, in the future we are planning to expand the platform contents' to handle other varieties, as well as specialized vocabulary (which may well also display significant regional variations). A good forward-thinking design is one that accommodates such variation and provides an easy path for content developers to add regional content and an easy way for interested students to explore it.Graph DBs play a crucial role in permitting the gradual addition and subsequent navigation of content. We will elaborate on this aspect of the design in the next section[9].

Toponyms and anthroponyms. The map of Morocco is full of names of Amazigh origin, referring to geological features (e.g. Ifrane), types of buildings (e.g. Agadir), plants and animals (e.g. Oualili), and tribe names (e.g., Aït Benhaddou), among others. Showcasing the presence of these names in the system, not only reinforces the cultural element of language learning, but also reinforces vocabulary in the head of learners, particularly those who have the opportunity to travel in Morocco. Therefore the Tamazight e-learning platform will make use of maps and geography of Morocco to help the learner navigate the language, the culture, the people and the country of Morocco.

4. Representing the Curriculumand Related Resources

In this section, we briefly review the curriculum that is planned for the Tamazight e-learning platform and its overall representation in a GDB. The section is divided into3 parts, discussing the structure of: 1) the course as whole and its core units, 2) Unit 0: Letters and Sounds, and 3) the dictionary.

9 To our understanding, IRCAM's standardization efforts have primarily impacted orthography,word segmentation when individual words tend to "run together", and variations in specific syntactic contexts, with a view to developing a consistent set of rules for the written language. With respect to vocabulary, given the unequivocal presence of regional variants, 'standard'acquires a different meaning. Choosing a specific word to refer to an object or action as 'standard' really means choosing a word that is likely to be understood widely, even if it is not the word of choice or the most frequent in a specific region to convey that meaning. That is, it is a choice that is likely to be more useful than others for the purposes of communication.

4.1.Curriculum Structure

The structure of the curriculum is based on the organization of materials in the AUI textbook, with some additions. At the top level the course node has edges to a *course* introduction, the body of the *course* consisting of ten *units*, a grammar compilation, a *dictionary*, and a stock of multimedia. The course introduction is pretty much self-standing, but the remainder of the components are heavily interlinked and either being a resource (the multimedia stock, the dictionary, and the grammar compilation) or drawing upon resources (the units).

Unit Structure. Each unit (there are ten in the current AUI textbook)has a similar organization. A given unit has properties number and title. It also has edges to a variety of nodes:

- The preconditions for entering the unit (if enforced), specified in terms of learning outcomes that should have been achieved;
- The intended learning outcomes (ILOs) of the unit, which function as the expected postconditions of the unit;
- The sounds of the unit, drawn from the letter-and-sound resources of Unit 0 (see Section 4.2);
- The dialogue(s) of the unit, exemplifying the communicative function on which the unit focuses;
- The word list of the unit, linked to the contents of the dictionary;
- Thegrammar of the unit, which is linked to the resources in the grammar compilation;
- The readings-and-writing resources associated with the unit; these might include:
    - Texts with corresponding audio(for reading practice) or without (for translation activities), and/or
    - Audio and video recordings, with their corresponding text script, for listening comprehension and dictation;
- The culture elements of the unit, which may be drawn from a variety of resources.

In addition, for the different activities there will be *formative assessments*, i.e. exercises that the learner goes through to assess their own learning (though instructors may inspect them too), and *evaluative assessments* which will determine whether the learner has achieved the ILOs of the unit and is ready to go on to the next unit (or level).

Multimedia Stock. The multimedia stock is a collection of resources that are called upon by the units for introducing new concepts or assessing student learning. There are several types of resources including dialogues, interviews, poetry, music audios and videos, documentary videos, stories, and images (drawings, photos, comic strips). All media types except images can include written text (e.g. lyrics of a song) and an audio/video component. Audios and video soundtracks together with text transcription can be used for learning pronunciation and listening comprehension. For reading practice, we are considering recording learners.

We have yet to design the grammar compilation and the elements it draws on as resources, but the structure of the dictionary is described in Section 4.3.

4.2. Letters and Sounds

Letters and sounds are the focus of Unit 0, the unit that precedes the AUI textbook and teaches learners the Tifinagh alphabet and its sounds. Earlier we described the planned sequence of presentation of letters and sounds; here we will describe the representation of these resources in the GDB implementation.Once again, the basic idea is to organize information in terms of a resource that can be called upon from different places in the curriculum. So, while a particular letter and its sound (or sounds) will be introduced in Unit 0, they can also be revisited and reinforced in later units in the curriculum.

A letter-and-sound resource, which is a node in the Graph DB includes the following elements:

- The written letter (the glyph), e.g. ⴹ;
- The name of the letter, e.g. ⵖⵄⴹ;
- A sound file giving its pronunciation in isolation;
- Written sound representations such as common conventions that use the extended Latin alphabet and are easy to remember, e.g.‘ḍ’, or “sounds as ‘d’ in ‘doh’ (preferred to using IPA)[10];
- Examples of words where the letter/sound is used, to be drawn from the dictionary and using preferentially words introduced in early units; and
- Occurrences of letter/consonant clusters where the letter sound is modified and corresponding examples of words and recordings, again preferentially drawn from the corpus of text provided in the two textbooks (an example is the cluster d‘d’+t‘t’, pronounced as ‘tt’).

4.3. Dictionary

The core principle in choosing a design for the dictionary is the desire to support the ability to represent lexical variation and even phonetic variation for the same lexeme. At the same time, we want the dictionary to represent the translation for an Amazigh word in more than one language, since the users of the system may come from different linguistic backgrounds. The Graph DB lends itself well to this task.

The concept encoded by a word is a node in the graph, call it NodeX. The edges linked to this node point to the realization of the conceptin different languages, including, for example, Modern Standard Arabic, Moroccan darija, English, and French, in addition to different lexical realizations in the Amazigh language. The latter may indicate regional differences (in which case they should be connected to the distinguishing geographical region on the map of Morocco) or different ways of expressing the same concept within a single region (in which case they should be connected to the same region)[11]. Similar reasoning can be applied to regional variations in pronunciation, which would be rendered by a sound file attached to the

10 Popularized by the character Homer Simpson in the TV show the Simpsons (there spelled d›oh). The OED has references from the BBC as far back as 1945, however. (Source: https://en.wiktionary.org/wiki/doh)

11 We will not consider here the situation in which the Amazigh word has a somewhat different meaning than the concept to which the word and its realizations in different languages are connected, although this situation is likely to arise and will need to be handled.

word and/or an alternative written representation as used in Unit 0. An example of this might be the pronunciation of the letter 'ⵖ', which is pronounced as the 'g' in 'girl' in the south but is realized differently, depending on its phonetic context, in the Middle Atlas.

5. Summary and Future Work

We have presented selected elements of the design of the Tamazight e-learning platform that is currently under construction at Al Akhawayn University in Ifrane. We have also offered justification for the technological choices we have made, and particularly the choice of a Graph DB to represent several aspects of the curriculum and content of the course. The general structure of significant components of the graph DB is in place, but

corrections to the AUI textbook content based on the input of IRCAM language specialists were recently completed and the interface for entering content into the system is still under development. We expect to have the interface completed by the end of the calendar year and to be entering content gradually as a way of testing the interface as it is being developed.

There are two sides to the web-based interface: what the learner interacts with and what the instructor(s) and developer(s) who place content in the system interact with. The latter actors have access to the learner's interface but not vice versa. With regard to the learners' interface, the main concern will be how to make the interface intuitive to use and easy to navigate to support flexible exploration without getting lost in the forest of paths defined by the graph representation. We have some ideas of how to use icons to do that and will actively be exploring these ideas soon. Concerning the instructor's interface, we are exploring the possibility of automatically generating at least a draft or skeleton of the interface, which can then be manually refined to achieve a more pleasing interaction. What we think will make it possible is Neo4j's ability to query its own structure using the Cypher query language to automatically generate information about the connections of a node, which can grow dynamically when more information is added to the DB.

References

Robinson, I., Webber, J., and Eifrem, E. (2013). *Graph Databases*. O'Reilly.

Bagui, B. & Ouakrime, M. (2011). *Tamazight Language Course, Book 1*. Al Akhawayn University in Ifrane.

La Constitution, Edition 2011. Royaume du Maroc, Secrétariat Général du Gouvernement (Direction de l'Imprimerie Officielle). Retrieved on June 30, 2016 from

http://www.amb-maroc.fr/constitution/ Nouvelle_Constitution_%20Maroc2011.pdf

Laraj, H. (2005). *Sawlat s tmazight* (Central Morocco). L'Institut Royal de la Culture Amazighe.

National Reading Panel (2000). *Teaching Children to Read: An Evidence-based Assessment of the Scientific Research Literature on Reading and its Implications for Reading Instruction*. National Institute of Child Health and Human Development.

Retrieved on June 30, 2016 from

https://www.nichd.nih.gov/publications/pubs/nrp/documents/report.pdf

Vers une démarche de qualité et de bonnes pratiques enseignantes des langues pour les filières scientifiques. Cas de l'intégration d'outils de simulation ludique et collaborative dédiés à l'apprentissage des langues naturelles

Aoued Boukelif
ICT's Research Team Communication Networks,
Architectures and Multimedia laboratory, Université Sidi Belabbas
aboukelif@yahoo.fr

Résumé

Cette communication vise à échanger les expériences et les bonnes pratiques enseignantes des langues afin d'asseoir un cadre de référence susceptible d'améliorer les compétences linguistiques et communicatives des étudiants scientifiques. L'objectif consiste à voir comment les TICE s'appliquent aux langues de grande diffusion déjà présentes sur le net et enfin d'évaluer les conditions requises en matière de démarches d'apprêtement linguistiques (syntaxe, lexique, phonétique, textes, graphies…) de tamazight afin de faciliter la diffusion de son enseignement sur le net.

Dans cette optique nous abordons l'intégration d'outils de simulation ludique et collaborative dédiés à l'apprentissage des langues. L'évaluation d'une séquence d'apprentissage met en lumière les obstacles à une bonne intégration de ces outils, et permet de mener une réflexion sur les remédiations à mettre en œuvre conduisant à terme à l'amélioration des pratiques enseignantes des professeurs de langue et terminologie intervenant dans les filières scientifiques.

1. Introduction

A l'ère du numérique, la technologie occupe une grande place dans notre vie. L'exploitation du numérique a des enjeux sociaux, culturels et économiques. Grâce au numérique, de nombreuses sociétés, sont devenues des sociétés du savoir et de l'information. Ces sociétés informatisées, grâce à la technologie, offrent aux citoyens des opportunités diverses dans les domaines de l'éducation, du commerce, de la santé, etc.

Les TICE offrent de nouvelles modalités d'enseignement / apprentissage, en particulier dans le domaine des langues . Grâce aux TICE en général et au multimédia en particulier, se développent de nouveaux dispositifs d'enseignement / apprentissage.

Dans le contexte de l'émergence des technologies nouvelles et face au rôle qu'elles semblent pouvoir jouer dans l'enseignement, c'est plus précisément l'intégration d'Internet dans un enseignement de langue qui est considérée ici. Un nouvel environnement techno-pédagogique est en train de s'imposer dans nos établissements scolaires. Notre objectif est de montrer que les apports des TICE sont nombreux et se manifestent fort lors de la mise en place de projets de classe motivants et valorisants, intégrés à des projets de communication authentique et que cette innovation interpelle des pratiques et des méthodes nouvelles exigeant aussi la prise en considération des aspects psychologiques et comportementaux.

Cet article porte sur l'intégration du numérique dans les pratiques pédagogiques des enseignants des langues. Il a pour objectif de présenter les principaux outils exploités dans l'enseignement-apprentissage des langues.

2. Didactique

Dans son sens classique, didactique est quasiment synonyme de pédagogique. Les dictionnaires indiquent que le mot vient du grec didaskein : enseigner, avec pour signification principale : «qui a pour objet d'instruire, qui est destiné à l'enseignement».

Si la didactique recherche bien un statut de type scientifique, elle ne relève pas d'une science exacte, qui permettrait de « prouver » comment marche l'apprentissage d'une notion, et d'obtenir à coup sûr le résultat espéré.

2.1. Triangle didactique

L'idée de triangle didactique a émergé simultanément à celle de triangle pédagogique (Houssaye).

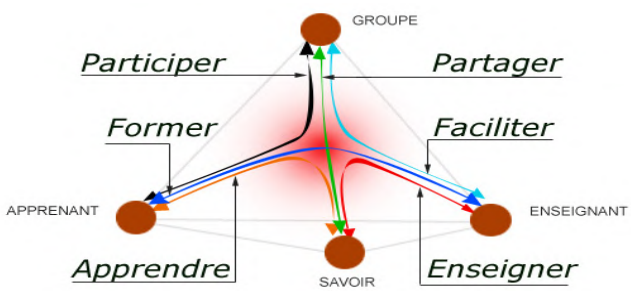


Figure 1. Tétraèdre pédagogique

Le caractère hétérogène des éléments du triangle a été souligné : une entité abstraite (le savoir), une personne identifiée (le formateur), mais un troisième élément aux contours imprécis, puisqu'on ne sait jamais exactement s'il s'agit de chaque apprenant individuel ou de l'ensemble du groupe en formation.

2.2. Quelle différence entre didactique et pédagogie ?

Le mot didactique est traditionnellement employé comme un adjectif, avec un sens assez voisin de celui de pédagogique. « *Qui a pour objet d'instruire* », « *qui a rapport avec l'enseignement*». Employé comme substantif, pour définir un nouveau champ de recherche. La didactique cherche dès lors à se démarquer de la pédagogie, celle-ci étant considérée comme obsolète, peu rigoureuse et souvent chargée d'idéologie.

L'accent est mis beaucoup plus nettement sur les contraintes de l'objet de savoir qui est l'enjeu de l'apprentissage. Didactique prend dès lors volontiers la marque du pluriel (les didactiques de disciplines).

La pédagogie pense la logique du savoir à partir de la logique de la classe. Elle se décline en pédagogies magistrale, non directive, différenciée, de groupes, par objectifs, etc., dont on voit bien que le contenu enseigné n'entre pas dans la définition. C'est la relation M-E (maître élèves, apprenant-formateur) qui est au premier plan et qui fonctionne comme variable indépendante, le savoir S étant introduit de façon seconde (peut-on faire de la pédagogie non directive en mathématiques ? de la pédagogie par objectifs en arts plastiques ? de la pédagogie différenciée en EPS ?

La didactique pense au contraire la logique de la classe à partir de la logique du savoir. Elle se décline en didactique des mathématiques, de la langue maternelle, des sciences, etc. C'est ici l'objet de savoir S qui est saillant, la question de la relation M-E passant au second plan sans être négligée. Les questions du mode d'intervention de l'enseignant, du travail de groupe ou de la nature du dispositif de travail, ne sont en effet pas absentes des préoccupations des didacticiens.

Didactique et pédagogie sont plus proches lorsqu'on les considère dans la pratique d'enseignement, qu'elles ne le sont en termes d'objet de recherche.

Les recherches en pédagogie portent principalement sur les courants pédagogiques et les grandes figures historiques de l'éducation, alors que les recherches en didactique se centrent sur l'analyse épistémologique d'un contenu, et sur les conditions d'efficience de son acquisition dans un cadre formatif.

3. Outils d'enseignement-apprentissage

3.1 Systèmes de gestion de contenu LMS, CMS, LCMS, DMS ,ECMS, Plateforme , Groupware, ENT , EIAH

Schéma de positionnement LCMS/LMS

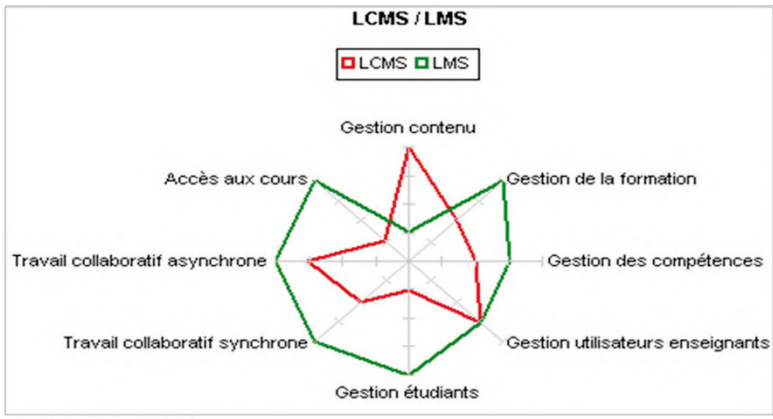


Figure 2. Schéma de positionnement LCMS/LMS

3.2. Cartes cognitives, cartes mentales et apprentissage incident de langue

Didactique des langues, Didactique de l'anglais, anglais précoce, TIC, simulation, jeu, MOO, anxiété technologique, anxiété linguistique, compétences, apprentissage incident ''incidental learning'', blogs, linguistique de l'énonciation ''utterer-centered linguistics'', jeux sérieux, jeux politisés ''social impact games'', fiction interactive, carte mentales ''mind maps'', pédagogie de l'enseignement d'une matière intégrée à une langue étrangère (EMILE). ''Content and Language Integrated Learning (CLIL)''.

- Online Collaborative Mind Mapping for English language learning.
- Tag clouds, keywords, Internet searching and incidental learning.
- Risk taking and linguistic pressure in written and oral production in ICT sessions.

3.3. Jeux sérieux

La plupart des définitions du terme «jeu sérieux» renvoient à un artefact informatique de type jeu vidéo dont l'usage vise à dépasser le simple divertissement pour atteindre des objectifs d'éducation ou de formation. Néanmoins, avec le développement des jeux sur dispositifs mobiles de type smartphone, les termes jeu vidéo ou jeu d'ordinateur ne rendent plus compte

totalement de la diversité des dispositifs. **L'expression «jeu sérieux» est alors employée pour nommer des supports et des technologies qui font appel au réseautage, à la réalité augmentée ou à la géolocalisation**. Le terme est aussi utilisé lorsque les composantes de la simulation sont intégrées au jeu vidéo afin de créer des environnements réalistes d'apprentissage gérés par des mécanismes de jeu et faisant souvent appel à la résolution de problèmes complexes.

D'autres auteurs utilisent le même terme pour décrire une situation d'apprentissage qui utilise des ressorts ludiques pour fonctionner. Ces situations sont des espaces de réflexivité au sein desquels le joueur/apprenant peut éprouver les stratégies qu'il élabore pour relever un défi. Les connaissances qu'il mobilise apparaissent alors comme des instruments permettant d'atteindre les objectifs fixés par le jeu. La conception de telles situations fait également largement appel aux technologies numériques d'aujourd'hui.

Au delà du choix des définitions, c'est, pour le chercheur, la portée d'un concept qui est en jeu. C'est aussi la question du périmètre de son objet d'étude qui est ainsi posée. Ce colloque souhaite accueillir des communications qui permettront de débattre de l'emploi de l'expression « jeu sérieux » dans différents travaux de recherche.

3.4. Drew argumentation

DREW est un outil Internet pour créer des situations d'apprentissage coopérant

DREW1 (Dialogical Reasoning Educational Web tool) est un environnement informatisé d'apprentissage humain collaboratif, développé en Java. Cet environnement est développé dans le cadre du projet européen SCALE2 (IST-1999) dont l'objectif pédagogique est de favoriser un apprentissage collaboratif basé sur l'argumentation (CABLE). En d'autres mots, l'objectif est d'amener les élèves du secondaire à « apprendre à travers des activités argumentatives ». Pour cela, à partir de l'environnement d'apprentissage DREW, différentes séquences d'enseignement et outils ont été conçus pour aider les élèves à acquérir, raffiner et étendre leur connaissance argumentative dans un domaine donné (le débat sur les OGM).

Exemple commenté de construction de diagramme argumentatif

SCALE est un outil intelligent pour l'apprentissage base sur l'argumentation (''Internet-based intelligent tool to Support Collaborative Argumentation-based Learning in secondary schools'' .

Illustration

Veuillez lire le débat suivant :

- Jules : Les voitures ne devraient pas être autorisées en centre-ville
- Julie : Vraiment ?
- Jules : Tu sais très bien qu'elles polluent !
- Julie : Oui, mais que veux-tu dire par «polluent» ?
- Julesanbsp;: Tu sais bien, la pollution de l'air par les gaz d'échappement des voitures ;

même le bruit est une forme de pollution

- Julie : Oui, c'est vrai que les voitures polluent ; tu devrais voir le nuage noir quand mon père démarre sa diesel.
- Jules : Oui, et c'est aussi dangereux pour les piétons
- Julie : Oui, mais tu ne peux pas empêcher les voitures d'aller dans le centre-ville
- Jules&bsp;: Vraiment ? Et pourquoi pas ?
- Julie : Les gens ont besoin de leur voiture dans le centre-ville
- Jules : Et pour quoi faire ?
- Julie : Pour faire leurs courses, pour aller au travail, tu vois ?
- Jules : Je ne suis pas d'accord, ils peuvent utiliser les transports en commun !
- Julie : Mais ce serait trop cher pour la ville d'avoir des bus pour tout le monde, et si on se débarrasse de toutes les voitures, les gens qui travaillent dans les usines perdront leur emploi
- Jules : Oui, je comprends, mais je ne l'accepte toujours pas. Parce que la pollution automobile est très mauvaise pour la santé
- Julie : Oui, après tout, je pense que tu as probablement raison, la santé des gens c'est très important, oui, même après ce que j'ai dit, ce n'est pas bon d'avoir des voitures en centre-ville.

Diagramme argumentatif

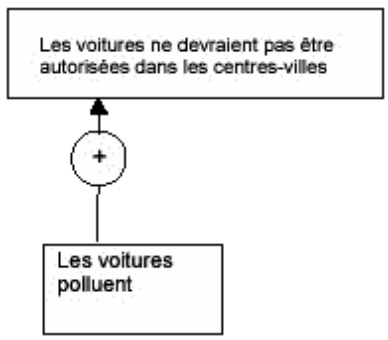


Fig. 3 - Diagramme argumentatif

Règle : Faites attention à la direction de vos flèches ! Si X est un argument pour ou contre Y, alors la flèche part de X et arrive à Y. En d'autres termes, elle va de X à Y.

Chaque cadre (argument) doit être relié au moins à un autre cadre ! Une proposition qui n'est reliée à rien n'est pas un argument.

3.5. Le e-Portfolio (ou e-Portfolio, portfolio numérique, portfolio électronique)

C'est une démarche visant à réfléchir sur ses projets (sociaux, professionnels) et définir quelles parties de ces projets communiquer à quel publics. Cette démarche peut être utilisée par les professionnels, les enseignants, les élèves, les parents ou autres personnes engagées dans une démarche de formation tout au long de la vie, dans le but de communiquer sur son profil ou garder des traces d'apprentissages.

Eportfolio d'apprentissage ou de développement: démarche qui démontre la progression et le développement des compétences de l'apprenant sur une période de temps. Ce dossier inclut en général des éléments d'auto-évaluation, des réflexions et des commentaires de tiers apportés par l'enseignant ou le tuteur permettant qu'une nouvelle forme de communication plus riche s'établisse entre eux.

L'e-portfolio est une nouvelle génération de CV. Modulable, interactif, il permettra de classer et de mettre à jour toutes ses connaissances et compétences, et de prouver ses réalisations professionnelles et sociales.

Qu'est-ce qu'un portfolio?

Il existe une multitude de définitions du terme « portfolio », de la plus usuelle : « Collection d'œuvres propre à refléter le talent de son auteur » à la plus éducationnelle : « Collection structurée des travaux d'un étudiant et des commentaires qui leur sont attachés, qui fait foi de ses compétences montrant des traces pertinentes de ses réalisations. Les travaux conservés, et parfois affichés, sont sélectionnés en fonction de critères établis par l'étudiant et par l'enseignant ».

Un portfolio numérique est un système de gestion informatique qui permet d'administrer un ou plusieurs portfolios, individuellement ou en groupe, peu importe le support : DVD, clé USB ou serveur télématique. Évidemment, chacun de ces supports offre ses avantages et ses inconvénients. Un système informatique de gestion centralisée est pratique, facile d'accès, accessible de partout si le système de protection des accès extérieurs n'est pas trop hermétique, mais il est parfois complexe, lourd à administrer, coûteux et souvent l'interopérabilité n'est pas assurée . Ainsi, l'étudiant qui quitte l'établissement ne peut apporter son portfolio et l'exploiter dans un autre établissement qui n'a pas exactement le même système de gestion des portfolios.

4. Conclusion

Cette communication s'inscrit dans l'émergence des nouvelles problématiques quant à la diffusion des connaissances (accessibilité du contenu numérique, nouveaux enjeux des systèmes de documentation, modalités d'appropriation, enseignement à distance, TICE, jeux électroniques éducatifs), l'accès à l'information. L'intégration de ces outils de simulation ludique et collaborative dédiés a l'apprentissage de la langue amazighe nous assistera à totalement repenser les espaces de formation tels que les espaces pédagogiques interactifs (EPI), et les plateformes d'enseignement numérique « e-cours » . Cette appropriation des outils (scenario d'apprentissage , cartes conceptuelles et exerciseurs pour l'apprentissage), services, et ressources numériques , contribuera a l'intégration des TICE dans les pratiques pédagogiques et a la conception et au développement de contenus pédagogiques multimédia interactifs dédiés à la langue amazighe tels que la structuration pédagogique modulaire d'un cours en ligne et la conduite d'un projet pédagogique.

Références

Abdelali A., Cowie J., Soliman H. S. (2005). Building a modern standard Arabic corpus. Actes du computational modeling of lexical acquisition workshop , pp. 1-7.

http://iportfolio.fr/

http://www.robertbibeau.ca/portfolio.html

http://scale.emse.fr/scale_tools/tutorial+examples/Training_Expe/pws/training_fr/index.html

http://tecfa.unige.ch/moo/tecfamoo.html

http://moo.echoduet.net/list.php

E-conte au service de l'apprentissage de l'amazighe

Siham Boulaknadel **Fadoua Ataa Allah**
Centre des Etudes Informatiques, Systèmes d'Information et de Communication, IRCAM
{boulaknadel, ataaallah}@ircam.ma

Résumé

Selon plusieurs études, la lecture joue un rôle crucial non seulement dans l'apprentissage des langues, mais aussi dans le développement global de l'enfant. Elle favorise le développement de la logique et du raisonnement, les aptitudes sociales et l'estime de soi, la transmission des informations scientifiques et culturelles.

Dans la perspective de donner le goût de lecture chez des jeunes apprenants de la langue amazighe à la fois à la maison et à l'école, un projet d'élaboration d'un site web de contes amazighes est en cours de réalisation.

Etant donné que les jeunes générations sont envahies par la technologie, ce projet a pour vocation de favoriser un environnement de lecture et de faire découvrir à un public enfant une sélection de contes variés, embrassant l'aspect culturel amazighe, pour une démarche pédagogique et ludique de promotion de la lecture et d'ouverture au monde.

Ces contes sont accompagnés d'une voix assurant la lecture dans une perspective de développer la concentration de l'enfant et son écoute, son lexique et son vocabulaire, son imagination et sa créativité, les notions de séquences dans le temps et son sens de l'observation.

1. Introduction

Le monde d'aujourd'hui connaît une extension de plus en plus grande pour la télévision et la technologie moderne, en dépit de la lecture qui demeure la plus riche source d'informations scientifiques et culturelles. Pourtant, elle représente une des techniques cognitives qui facilite la compréhension, enrichit l'expérience du lecteur et participe au développement de sa personnalité. Cependant, le conte reste toujours un moyen attractif pour les enfants, un support idéal sur lequel peuvent s'appuyer les enseignants pour transmettre les savoirs fondamentaux, en particulier la maîtrise de la langue orale et écrite. Il véhicule des valeurs sociales et civiques. D'ailleurs, selon Bruno Duborgel « Le conte permet de faire le lien entre les différentes disciplines scolaires guidant l'enfant vers un savoir pluriel »[1].

Face à des enfants qui évoluent depuis leur naissance dans une société irriguée par le numérique, les manières d'apprendre et d'enseigner doivent être profondément repensées à base des transformations des modes de production et de diffusion de l'information et des

1 Ecrit de Bruno Duborgel (Duborgel, 1992).

connaissances. Ainsi, ce projet, d'élaboration d'un site Web pour les contes amazighes, est né de la volonté d'exploiter les usages personnels que les enfants ont de la technologie pour les mettre au service de l'apprentissage, notamment de la langue amazighe. Il vise de faire du conte numérique un levier pour aider les enfants à mieux lire et s'exprimer, pour les aider à développer leur pensée et leur jugement esthétique. Plutôt que d'opposer la culture numérique des enfants aux savoirs, conjuguons-les pour encourager les enfants à devenir des acteurs éclairés d'une véritable culture littéraire et numérique.

Le reste de cet article est composé d'une section sur le rôle du conte dans le processus d'apprentissage des langues, notamment les langues maternelles ; suivie d'un descriptif dudit projet, en illustrant sa conception ainsi que ses modules. En outre, la dernière section conclu le travail et décline une liste de perspectives présentant les étapes à entreprendre.

2. L'apprentissage des langues via le conte

Le conte est un support incontournable. Souvent, il fait partie des œuvres d'art qui appartiennent au patrimoine culturel de l'humanité et qui représentent la vision du monde, contribuant ainsi à la transmission d'une culture commune, véhiculant des valeurs, des comportements rituels et des règles d'organisation sociales. De ce fait, le conte provoque l'enthousiasme des enfants et les aide à se familiariser avec des formes linguistiques et stylistiques nouvelles. Il leur permet de s'évader du quotidien, de développer leur imagination et de s'ouvrir aux différentes cultures en illustrant la richesse de la culture universelle à travers les différences.

Autrefois, le conte était caractérisé par la seule transmission orale. Il prendrait ses sources il y a cinq-cent-mille ans après la naissance du feu. Les hommes se regroupaient autour des foyers pour se protéger du froid, des bêtes mais également pour parler. Et, pour donner une explication à ce qu'ils ne connaissaient pas, ils se sont mis à supposer, à rêver, à inventer. Les hommes ont essayé d'expliquer les phénomènes de la nature qui les intriguaient : l'arc en ciel, la lune, le soleil, les éclipses, ... (Renoux, 1999). Ceux-ci se transmettaient par la bouche à oreille.

La structure répétitive de ces récits facilite leur transmission, mais face à la migration rurale et l'industrialisation, la préservation de cette culture orale à travers des écrits s'est développée surtout dès le 17ème siècle. Dès lors, les contes existent sous une forme écrite, qui essaient de donner une transcription fidèle à l'oral. Ce qui a permis d'en faire objet d'études scientifiques, notamment en littérature (Propp, 1928) et en psychanalytique (Bettelheim, 1995).

Aujourd'hui, la pratique du contage a quasiment disparue. Cependant, le conte est devenu un genre littéraire. Il est véhiculé par différents types de supports : l'imprimé, le film, le théâtre et le dessin animé. Généralement, les contes possèdent une structure spécifique avec des variables et des constantes (Propp, 1928). Ils peuvent être regroupés en plusieurs catégories, qui varient selon les auteurs et les ouvrages. Cependant, la classification la plus répandue est celle d'Aarne-Thomson[2], dont nous citons :

- **Les contes d'animaux** : ce genre de conte met en scène uniquement des animaux doués de parole et se comportant comme des humains.

2 La classification internationale Aarne-Thomson est le résultat du travail du finlandais Antti Aarne (Antti Aarne, 1961), et l›américain Stith Thompson.

- **Les contes étiologiques** : des récits qui expliquent le pourquoi et le comment des choses de la réalité, en y répondant d'une manière fantaisiste.
- **Les contes facétieux** : ce type de conte regroupe toutes sortes de récits différents, souvent anecdotiques. Il présente des anti-héros ayant échoué sous la forme d'anecdotes, dont la plupart sont destinées aux adultes, seules certaines histoires d'idiots sont adaptées aux enfants.
- **Les contes formulaires** : ou randonnées ou en chaîne se distinguent par la fixité de leur forme. Ils se basent sur des pirouettes verbales par lesquelles le conteur signifie qu'il ne veut rien dire, soit qu'il amène l'auditeur à poser une question à laquelle il répond par une moquerie.
- **Les contes merveilleux** : ces contes tiennent une place très importante dans la littérature orale. Ils mettent en scène des personnages évoluant dans un monde magique où des personnages secourables aident les héros à accomplir les épreuves et à triompher des obstacles.
- **Les contes nouvelles** : certains de ces récits se trouvent entre le conte et la nouvelle. Ils présentent l'intelligence, le courage et l'astuce à travers des personnages prenant en main leur destin contrairement aux héros stéréotypés des contes merveilleux qui sont quasiment mus de l'extérieur.
- **Les contes de l'ogre dupé** : ces récits racontent les aventures d'un garçon ou d'un homme futé qui, par son astuce et sa persévérance, se joue de la méchanceté et de la bêtise de l'autre.

A travers toutes ces variétés, le conte « tout en divertissant l'enfant, l'éclaire sur lui-même et favorise le développement de sa personnalité. Il a tant de significations à des niveaux différents et enrichit tellement la vie de l'enfant qu'aucun autre livre ne peut l'égaler »[3].

Etant donné l'importance de la lecture dans l'apprentissage des langues en général, et les langues maternelles en particulier (Miller, 2008), notamment dans les domaines de la lecture et de l'expression écrite, et de tout ce qui précède, le conte devrait occuper une place importante dans le processus de l'apprentissage des enfants, à la fois à l'école et à la maison. Ainsi, au fils des cycles scolaires, l'enfant sera mis en situation de découvrir le plaisir du conte. Il aura l'occasion d'enrichir les échanges et le langage d'évocation, d'ouvrir son esprit à la culture des contes et des légendes dont les significations sont universelles, s'essayer à un usage du langage plus complexe. A l'aide de ses enseignants et parents, l'enfant s'initiera à construire le langage d'évocation des évènements passés ou à venir, le langage de communication et d'échange de connaissances ainsi que le langage de l'argumentation (Mongin, 2006). Ce qui conditionne la réussite des apprentissages ultérieurs.

3 Ecrit de Bruno Bettelheim (Bettelheim, 1995).

3. Description du projet

Partant de l'idée que les contes sont de formidables vecteurs d'apprentissage des langues, nous avons entrepris, au sein du Centre des Etudes Informatiques, des Systèmes d'Information et de Communication (CEISIC), l'élaboration d'une plateforme de contes numériques amazighes pour mettre à disposition des enseignants et des parents de nombreuses ressources pédagogiques et idées pour mener des cours, ateliers et activités autour des contes.

La plateforme est conçue dans une vision globale qui prévoit, dans une démarche progressive, l'élaboration d'une plateforme à raconter des contes pour enfants fascinante et instructive, qui facilitera à travers la voix, textes et images l'apprentissage de la lecture et la compréhension du texte.

Ainsi, et par l'intermédiaire des nouvelles technologies qui facilitent la structuration des données, l'enfant désireux d'apprendre trouve à sa disposition une plateforme multimédia interactive simple et facile à utiliser.

3.1. Conception de la plateforme des contes numériques

Au cours de l'élaboration de la plateforme, plusieurs critères ont été pris en compte en matière d'enrichissement, d'adaptation à l'esprit des enfants et de contribution à la consolidation de leur apprentissage, notamment :

- doter la plateforme d'une base de données dynamique, offrant la possibilité d'enrichir la plateforme de façon continuelle ;
- adopter une structure composée du conte ; l'illustration utilisée comme moyen d'attractivité et la voix comme outil d'immersion ;
- structurer les contes de la plateforme selon trois catégories. Le premier concerne l'usage des contes dans les pièces théâtrales, le deuxième comporte les contes merveilleux et le troisième comprend les contes animaliers ;
- enrichir la plateforme d'un ensemble de comptines et de jeux ludiques pour permettre à la fois attirer l'attention de l'enfant, ancrer son apprentissage et évaluer ses acquis.

Afin de réaliser ces perspectives, une conception générale de la plateforme a été réalisée (cf. Figure 3.1), identifiant l'ensemble des acteurs et déterminant leurs rôles.

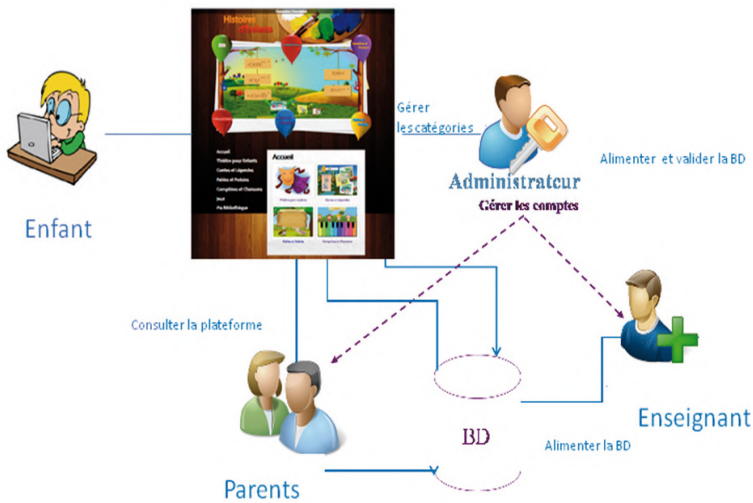


Figure 1 : Structure de la plateforme

L'utilisateur, généralement est un enfant ou un parent, a la possibilité de :

- Consulter le contenu de la plateforme en sélectionnant l'une des catégories précitées.
- Jouer pour se divertir tout en évaluant les acquis et consolidant l'apprentissage à la lecture. Les activités proposées varient entre jeu de constitution de l'histoire du conte et jeu de questions-réponses pour évaluer l'apprentissage du conte.

L'enseignant dont la tâche consiste à la consultation en plus de l'alimentation du contenu de la plateforme qui ne sera pris en compte qu'après la validation de l'administrateur.

L'administrateur a un rôle crucial et pionnier dans l'organisation de la plateforme. Il a le droit de fixer les catégories des contes sur la plateforme, de créer des comptes pour les utilisateurs et il a la tâche d'alimentation du contenu de la plateforme.

3.2. Structure de la plateforme

Le projet qui est en cours de réalisation s'inscrit dans le cadre de l'accompagnement personnalisé en classe ou à la maison. La lecture sur support numérique est l'occasion de soutenir et d'approfondir les compétences de lecture des enfants, et de travailler les compétences méthodologiques comme l'expression orale. Ainsi, la plateforme (Figure 3.2) est conçue à partir de recueil de contes populaires en amazighe destiné aux enfants. Il a été opté, dans une première étape, une structuration comprenant quatre rubriques, à savoir :

- Les pièces théâtrales à base de contes sont présentées sur cette plateforme pour faire travailler la compréhension orale des apprenants et leur faire découvrir des œuvres théâtrales amazighes variées à travers une activité originale, ludique et motivante.
- Contes merveilleux présentent une sélection de contes inédits, suscitant rêve et bonheur.

- Contes animaliers proposent un florilège d'un patrimoine méconnu ou oublié, où les animaux jouent des rôles principaux.
- Comptines offrent un ensemble de chansons dans l'objectif de préparer les enfants à la lecture grâce à la mélodie qui les aide à capter le sens d'un message, avant même d'en comprendre le sens des mots.

Figure 2 : Page d'accueil de la plateforme

4. Conclusion

Dans cet article, nous avons présenté un projet en cours qui vise la diffusion et la démocratisation de l'accès à la lecture pour les enfants. L'objectif est de donner aux enfants, aux parents et aux enseignants l'accès à une collection riche de contes et comptines populaires amazighes. Dans la perspective de compléter ces contributions, nous avons prévu l'enrichissement de la base de données, le développement du module de jeu pour l'évaluation des acquis et pour rendre la lecture plus attractive, nous avons pensé de doter les contes d'un mode «karaoké» qui permet à l'enfant de suivre la lecture du texte.

Références

Antti A. (1961). *The Types of the Folktale : A Classification and Bibliography,* The Finnish Academy of Science and Letters, Helsinki,

Bettelheim B. (1995). Psychanalyse des contes de fées in «Parents Enfants», Robert Laffont, coll Bouquins.

Duborgel B. (1992). *Imaginaire et pédagogie*, Privat.

Miller S. (2008). *The Power of Story: Using Storytelling to Improve Literacy Learning*. Journal of Cross-Disciplinary Perspectives in Education Vol. 1, No. 1, pp 36 – 43.s

Mongin E. (2006). *Le conte : du projet d'apprentissage au projet culturel*. Institut Universitaire de Formation des Maîtres de Bourgogne.

Prop V. (1928). *Morphologie du conte*.

Renoux J. C. (1999). *Parole de conteur : Essai sur la pratique, l'historique et les approches du conte,* Edisud, coll L'espace du conte.

Perception évolutive envers les Serious Games et à la création vidéoludique par des enseignants stagiaires de la langue amazighe

Ibrahim Ouahbi[1,2] Hassane Darhmaoui[2] Fatiha Kaddari[1]
[1] Laboratoire de Didactique, d'Innovation Pédagogique et Curriculaire (LADIPEC), Faculté des sciences Dhar Elmahraz, Université sidi Mohammed Ben Abdellah, Fès Maroc
[2] Center for Learning Technologies (CLT), Al Akhawayn University, Ifrane, Morocco
ibrahim.ouahbi@usmba.ac.ma, h.darhmaoui@aui.ma, kaddari@yahoo.fr

Résumé

Nous avons mené une expérimentation comportant deux types d'activités auprès de 20 enseignants stagiaires de la filière langue Amazighe au Centre Régional des Métiers de l'Education et de la Formation (CRMEF) de la ville de Nador. La première activité est basée sur la pratique des «*Serious Games*» et l'autre sur la création de jeux simples à l'aide de l'environnement Scratch qui est un environnement open source facilitant la création vidéoludique et connu pour sa vocation pédagogique. En amont et en aval de l'expérimentation, un même questionnaire a été distribué aux mêmes enseignants stagiaires pour mesurer leurs perceptions envers les jeux vidéo éducatifs. Les résultats de l'étude comparative de l'évolution de la perception envers les «*Serious Games*» ont montré une évolution considérable et une bonne prise de conscience sur le potentiel de ces jeux en classe. La totalité des participants ont apprécié l'environnement Scratch et son potentiel pédagogique qui permet de promouvoir l'apprentissage de la langue amazighe et l'acquisition des compétences transversales. Ils ont aussi éprouvé un besoin d'apprendre plus sur l'utilisation et la création des «*Serious Games*» en classe.

1. Introduction

Récemment les jeux vidéo sont devenus les activités de loisirs les plus sollicitées par les jeunes (Boyle *et al.* 2012). Plusieurs recherches issues de plusieurs disciplines et en particulier du domaine de l'éducation appellent à l'exploitation de la puissance cognitive des jeux vidéo pour mieux engager et motiver les élèves dans leurs apprentissages (Gee 2011; Granic *et al.* 2014).

Le terme «*Serious Game*» est communément utilisé pour désigner les jeux vidéo dont la finalité première est autre que le simple divertissement (Michael *et al.* 2005). Ce qui différencie un Serious Game d'un jeu vidéo de divertissement est que les Serious Games sont explicitement et intentionnellement conçus à des fins sérieuses, pourtant, rien n'empêche d'utiliser à la place les jeux vidéo de divertissement aux mêmes finalités. En effet on peut procéder à un

détournement des règles premières d'un jeu vidéo classique pour en ressortir des applications utiles. Cet usage des jeux vidéo de divertissement vers une finalité utilitaire est désigné par le terme «*Serious Gaming*» (Kasbi 2012). Les ouvrages de Gee (2007), Shaffer (2006) et Presky (2007) présentent plusieurs exemples de jeux vidéo de divertissement qui peuvent être bénéfiques dans le domaine éducatif et de la formation.

Selon Marfisi-Schottman (2013) plusieurs termes sont utilisés pour désigner le terme Serious Game, qui diffèrent selon le contexte d'utilisation. Sawyer et Smith en recensent plusieurs termes représentant le vocal Serious Game : «*Educational Games, Simulation, Virtual Reality, Alternative Purpose Games, Edutainment, Digital Game-Based Learning, Immersive Learning Simulations, Social Impact Games, Persuasive Games, Games for Change, Games for Good, Synthetic Learning Environments, Games based 'X'…*» (sawyer *et al.* 2008).

Dans notre travail nous désignons par Serious Game toute application informatique utilisant des ressorts ludiques issus du jeu vidéo et dont l'objectif est de faciliter l'apprentissage et l'enseignement. L'accent a été mis sur l'étude de l'évolution de la perception des enseignants stagiaires marocains envers les Serious Games. Nous tentons répondre à la question de recherche suivante :

«*L'introduction des Serious Games dans la formation des enseignants, entraine-elle un changement positif dans les représentations et les attitudes des futurs enseignants envers cette innovation pédagogique ?*»

En effet, dans ce contexte, il nous semble évident que cette étude de perception est essentielle pour toute intégration future de ces jeux dans le système éducatif marocain. Nous croyons qu'une bonne conscience des enseignants du potentiel des Serious Games entraînera certainement leur acceptation comme outils pédagogiques innovants en classe. Ainsi nous avons dans un premier temps dressé un état de l'art des travaux rattachés à l'étude de la perception et à la sensibilisation des enseignants envers les Serious Games. Ensuite, nous avons réalisé pendant 3 séances (2 Heures / séance) des activités autour de l'utilisation des Serious Games, ainsi que la création de jeux simples à l'aide de l'environnement Scratch (Maloney *et al.* 2010) avec 20 enseignants stagiaires de la filière «Langue Amazighe» au CRMEF de la ville de Nador. En amont et en aval de l'expérimentation, un même questionnaire a été distribué aux mêmes enseignants stagiaires pour mesurer leurs représentations envers les Serious Games. L'étude comparative de l'évolution de la perception a montré une très bonne conscience sur le potentiel et les possibilités d'utiliser ces jeux en classe.

Durant les discussions menées avec les enseignants stagiaires, la majorité d'entre eux envisage utiliser les jeux dans leurs pratiques enseignantes. Ils ont exprimé leurs besoins en formation sur l'intégration des jeux en classe. Ils recommandent l'utilisation d'environnements facilitant la réalisation des Serious Games comme Scratch dans la formation aux CRMEF ou dans la formation continue des enseignants.

2. Etat de l'art

L'adoption et l'efficacité de l'apprentissage basé sur les jeux vidéo dépend en grande partie de l'acceptation des jeux vidéo par les enseignants comme moyen et outil didactique dans leurs pratiques enseignantes (Can *et al.* 2006; Bourgonjon *et al.* 2013). En effet les attitudes des professeurs envers les jeux vidéo acquièrent une grande importance parce que ce sont eux qui vont utiliser les jeux en classe.

Plusieurs recherches se sont intéressées à l'étude de la perception des enseignants avant et après leurs expositions à des séances de sensibilisations sur les Serious Games. Dans une expérimentation menée par Baker et al (2006) ; 49 professeurs en cours de formation ont joué à 3 types de jeux : des jeux de puzzle «puzzle game», des jeux de stratégies «strategy game» et des jeux de tir à la première personne «First Person Shooter - FPS», pour ensuite les inviter à analyser la façon d'utiliser ces jeux en classe, ainsi que d'explorer les différentes possibilités offertes par ces jeux. Les résultats de cette étude montrent que les enseignants peuvent utiliser certains jeux commerciaux en classe, et soulignent que l'enseignant doit être attentif lors de la sélection du jeu à utiliser. Les résultats de l'expérimentation menée par Romero *et al.* (2015) avec 51 enseignants stagiaires ont révélé qu'avant l'expérimentation ces étudiants ignoraient l'existence des Serious Games, ainsi que les possibilités de réutiliser les jeux existants ou en créer de nouveaux en utilisant les environnements de création de jeux. Après des activités basées sur les jeux, la plupart de ces futurs enseignants ont réussi à identifier et intégrer les serious games dans les curriculums scolaires. Pourtant cette étude a montré que ces futurs enseignants évitent les activités basées sur l'adaptation, la personnalisation ou la création de nouveaux jeux.

Au Maroc, peu de recherches se sont intéressées à l'étude de la perception des enseignants envers les Serious Games. Les résultats de notre expérimentation auprès de 20 stagiaires de la filière langue Amazighe (Ouahbi *et al.* 2014a) ont indiqué que la plupart des futurs enseignants, bien qu'ils aient pratiqué le jeu vidéo («gamer generation») ; ignorent quelques aspects d'usage du potentiel de ces ressources dans l'enseignement. Une fois qu'ils ont été initiés aux jeux éducatifs, ceux d'entre eux qui pensent utiliser les jeux éducatifs dans leurs pratiques enseignantes a doublé de 40% à 80%. Néanmoins, ces participants ont soulevé la problématique de la rareté des Serious Games pour l'apprentissage des langues amazighe et arabe. Même si elles existent, ces ressources ne répondent pas parfaitement aux besoins didactiques nécessaires à enseigner quelques concepts. Ils ont aussi souligné qu'ils ignoraient les possibilités de création des jeux vidéo.

Il est clair que la personnalisation et la création de jeux demandent des compétences informatiques avancées (Earp *et al.* 2013), pourtant des environnements informatiques ont été développés pour faciliter la programmation aux débutants à travers la création des jeux, d'animations et d'histoires (Ouahbi *et al.* 2015b). Dans ce contexte, plusieurs travaux ont prouvé l'efficacité de ces environnements, et en particulier l'environnement Scratch (Maloney *et al.* 2010) dans l'enseignement en plus des concepts de base de programmation : l'acquisition des compétences transversales comme la communication ; le partage ; la pensée critique ; la résolution de problèmes ; l'autonomie et la créativité chez les élèves (Ouahbi *et*

al. 2014b ; 2015a). L'environnement Scratch[1] a fait ses preuves aussi au niveau universitaire. Kim et al (2012) l'ont utilisé dans des cours de programmation au profit d'enseignants en cours de formation. Ces étudiants ont créé des animations, des histoires et jeux à des fins éducatifs. Ils ont développé des compétences en programmation d'une façon implicite et ils ont renforcé leurs capacités en TIC pour être capable d'innover et d'intégrer les TIC dans leurs pratiques enseignantes. De même Baron *et al.* (2013) ont mené une expérimentation d'initiation à la programmation en utilisant l'environnement Scratch auprès d'étudiants de Master de sciences de l'éducation dans le cadre d'une pédagogie par projet. Les résultats de cette étude montrent qu'il est possible, de faire produire des programmes relativement intéressants par des personnes non formées auparavant en informatique.

Parmi les principaux obstacles dans l'implémentation des jeux en classe, figurent: le manque de souplesse dans les programmes d'enseignement, les effets négatifs des jeux, le manque de supports et de connaissances pour intégrer efficacement les jeux vidéo en classe (Becker *et al.* 2005 ; Baek 2008 ; Hsu *et al.* 2011). Néanmoins, la plupart des recherches ci-dessus agréent que les enseignants ont des perceptions positives concernant l'utilisation des jeux vidéo en classe, et croient que les jeux vidéo éducatifs seront largement utilisés dans l'avenir.

3. Méthodologie

3.1. Déroulement de l'expérimentation

3.1.1. Etude de la perception de notre échantillon envers les Serious Games

Notre échantillon est constitué de 20 professeurs stagiaires de la filière «Langue Amazighe» qui suivent leur formation au CRMEF de la ville de Nador.

Nous avons commencé notre expérimentation par l'identification des représentations de ces stagiaires envers les Serious Games. Ils ont été appelé à répondre à un questionnaire «pré-enquête» sur leurs perceptions. Ensuite, Ils ont suivi une formation de 6 heures étalées sur 3 séances. Nous avons organisé deux types d'activités : l'une basée sur le Serious Games et l'autre sur la création des jeux simples.

3.1.2. Activités basées sur la pratique des Serious Games

Durant une première séance de 2 heures ; les 20 enseignants stagiaires ont été amenés à jouer à des Serious Games tels que le jeu «September 12th»[2], 2025 exmachina[3] et d'autres jeux à caractère éducatif par exemples les jeux disponibles sur le site de l'Institut Royal de la Culture Amazighe IRCAM[4]. L'objectif était de faire savoir aux étudiants le concept de «*Serious Game*» ainsi que découvrir des Serious Games dans l'enseignement des langues.

1 https://scratch.mit.edu/
2 http://www.gamesforchange.org/play/september-12th-a-toy-world/
3 http://www.2025exmachina.net/
4 http://www.ircam.ma/fr/index.php

3.1.3. Activités basées sur la création des Serious Games

Durant une 2ème séance de 2 heures, les enseignants stagiaires ont été initiés à l'environnement Scratch en réalisant des petits jeux, histoires en langue Amazighe et des animations accessibles à partir du menu «Aide» du logiciel. Au cours de la 3ème séance les stagiaires ont été amenés à créer des Serious Games en utilisant l'environnement Scratch. L'objectif était de les initier à créer leurs propres jeux. A la fin de cette séance, nous leurs avons proposé de réaliser des fiches pédagogiques d'une leçon qui porte sur l'utilisation d'un Serious Games et qui sera en relation avec l'enseignement de la langue amazighe.

A noter que pendant nos 3 séances, nous avons mené des discussions avec les stagiaires sur les différentes manières et approches pour une meilleure exploitation de ces jeux en classe.

3.2. Etude de l'évolution de la perception: post-enquête

A la fin des séances de formation sur les Serious Games, les enseignants stagiaires ont répondu à un deuxième questionnaire pour mesurer l'évolution de leur perception envers les jeux vidéo supportant l'apprentissage.

4. Résultats et discussions

4.1. Résultats de l'étude de la perception

4.1.1. Résultats de la pré-enquête

L'analyse des résultats de la pré-enquête envers la perception des 20 enseignants stagiaires a montré que la totalité des enseignants stagiaires ignorent l'existence du terme Serious Game et pensent qu'ils n'ont pas le savoir nécessaire pour contribuer à la création d'un jeu vidéo éducatif. Ces résultats corroborent avec celles de Romero et al (2015).

Environ 20% des stagiaires pensent être capable de sélectionner et d'utiliser des jeux vidéo se rapportant à la langue amazighe. Néanmoins plus de 75% d'eux pensent que l'intégration des Serious Games en classe peut développer l'imagination et motiver les élèves dans leurs apprentissages.

4.1.2. Résultat de l'étude de l'évolution de la perception

L'étude de l'évolution de la perception des 20 enseignants stagiaires envers les jeux vidéo comme outil didactique est résumée dans le tableau 1 suivant:

Question	Avant	Après
Je crois que les jeux vidéo éducatifs peuvent être utilisés en classe pour :		
• Aider les élèves à développer leur imagination	80%	95%
• Développer leurs compétences en communication	40%	95%
• Améliorer la capacité de la résolution de problèmes chez les élèves	45%	85%
• Apprendre à coopérer avec les autres	40%	80%
• Motiver les élèves pour l'apprentissage	70%	100%
• Aider les élèves dans leur apprentissage à la maison	40%	60%
• Aider les élèves dans leur apprentissage en classe	40%	85%
• Faciliter le travail des enseignants	40%	85%
Je pense que je suis capable de contribuer à la création d'un jeu vidéo éducatif	5%	55%
Je pense que je suis capable de sélectionner un jeu vidéo se rapportant à la matière que je vais enseigner	45%	95%
Je pense utiliser les jeux vidéo éducatifs dans mes pratiques enseignantes	45%	85%

Tableau 1: Étude comparative de l'évolution de la perception envers les jeux vidéo éducatifs avant et après l'expérimentation

4.2. Discussion

Les séances menées ont donné l'occasion aux 20 enseignants stagiaires de la filière «langue Amazighe » qui ont participé à notre expérimentation d'apprécier l'apport éducatif que peuvent apporter les Serious Games en classe. Ces séances ont aussi permis aux futurs enseignants d'être conscient du potentiel des jeux vidéo pour l'apprentissage, et de penser à innover dans leurs pratiques enseignantes. Les stagiaires ont apprécié l'environnement Scratch et ont pu saisir son potentiel éducatif. Leurs réalisations avec ce logiciel étaient très simples, cela est dû à notre avis au nombre réduit des séances réalisées vu leurs agendas chargés pendant leur formation professionnelle. A la fin, ils recommandent l'utilisation d'un environnement facilitant la création vidéoludique comme Scratch dans la formation aux CRMEF ou dans la formation continue des enseignants.

5. Conclusion

Les résultats de notre travail ont indiqué que la plupart des futurs enseignants ignorent l'existence des Serious Games ainsi que quelques aspects de l'usage du potentiel de ces ressources dans l'enseignement de la langue amazighe. Pourtant une fois qu'ils ont été initiés aux Serious Games, leur perception a considérablement évolué d'une moyenne de 45% à 84%. Les stagiaires ont apprécié l'environnement Scratch et ont ressenti son potentiel éducatif. Ils recommandent l'utilisation d'un tel environnement facilitant la création vidéoludique dans la formation aux CRMEF ou dans la formation continue des enseignants. Ils pensent que cette innovation pédagogique ne peut être efficace que si elle est appuyée par le ministère de tutelle. Au-delà des problèmes d'ordre matériels, les stagiaires soulèvent le problème de la perception négative des parents et de l'administration aux jeux vidéo comme un support éducatif.

Références

Baek, Y. K. (2008). What hinders teachers in using computer and video games in the classroom? Exploring factors inhibiting the uptake of computer and video games. CyberPsychology & Behavior, 11(6), 665-671.

Bakar, A., Inal, Y., & Cagiltay, K. (2006). Use of Commercial Games for Educational Purposes: Will Today's Teacher Candidates Use them in the Future?. World Conference on Educational Multimedia, Hypermedia and Telecommunications, 2006(1), 1757–1762.

Baron, G. L., & Voulgre, E. (2013). Initier à la programmation des étudiants de master de sciences de l'éducation? Un compte rendu d'expérience. In Sciences et technologies de l'information et de la communication (STIC) en milieu éducatif.

Becker, K., & Jacobsen, M. (2005). Games for Learning: Are Schools Ready for What's to Come?. Paper presented at the DiGRA 2005 Conference: Changing Views – Worlds in Play, Vancouver.

Bourgonjon, J., De Grove, F., De Smet, C., Van Looy, J., Soetaert, R., & Valcke, M. (2013). Acceptance of game-based learning by secondary school teachers. Computers & Education, 67, 21-35.

Boyle, E. A., Connolly, T. M., Hainey, T., & Boyle, J. M. (2012). Engagement in digital entertainment games: A systematic review. Computers in Human Behavior, 28(3), 771-780.

Can, G., & Cagiltay, K. (2006). Turkish prospective teachers' perceptions regarding the use of computer games with educational features. Educational Technology & Society, 9(1), 308-321.

Earp, J., Dagnino, F., Kiili, K., Kiili, C., Tuomi, P., & Whitton, N. (2013). Learner Collaboration in Digital Game Making: An Emerging Trend. Learning & Teaching with Media & Technology, 439.

Gee, J. P. (2007). What Video Games Have to Teach Us About Learning and Literacy. Palgrave Macmillan.

Gee, J. P. (2011). Reflections on empirical evidence on games and learning. Computer games and instruction, 223-232.

Granic, I., Lobel, A., & Engels, R. C. (2014). The benefits of playing video games. American Psychologist, 69(1), 66.

Hsu, T. Y., & Chiou, G. F. (2011). Preservice teachers' awareness of digital game-supported learning. Society for Information Technology & Teacher Education International Conference 2011(1), 2135-2141.

Kasbi, Y. (2012). Les serious games: une revolution. Edipro.

Kim, H., Choi, H., Han, J., & So, H. J. (2012). Enhancing teachers' ICT capacity for the 21st century learning environment: Three cases of teacher education in Korea. Australasian Journal of Educational Technology, 28(6), 965-982.

Michael, D. R., & Chen, S. L. (2005). Serious games: Games that educate, train, and inform. Muska & Lipman / Premier-Trade.

Maloney, J., Resnick, M., Rusk, N., Silverman, B., & Eastmond, E. (2010). The scratch programming language and environment. ACM Transactions on Computing Education (TOCE), 10(4), 1-15.

Marfisi-Schottman, I. (2013). Méthodologie, modèles et outils pour la conception de Learning Games (Doctoral dissertation, INSA de Lyon).

Ouahbi, I., Darhmaoui, H., Kaddari, F., Elachqar, A., Regragui, M., & Lahmine, S. (2014a). La connaissance et la perception des jeux vidéo comme outil didactique par les enseignants stagiaires de la langue Amazighe. A la 6ème édition de la conférence internationale sur les Technologies d'Information et de Communication pour l'Amazighe, Rabat, Maroc. http://tal.ircam.ma/conference/data/papers/27.doc

Ouahbi, I., Kaddari, F., Darhmaoui, H., Elachqar, A., & Lahmine, S. (2014b). Serious Games for teaching combined basic programming and English communication for non-science major students. International Journal on Advances in Education Research, 1(1), 77–89.

Ouahbi, I., Kaddari, F., Darhmaoui, H., Elachqar, A., & Lahmine, S. (2015a). Learning Basic Programming Concepts by Creating Games with Scratch Programming Environment. Procedia - Social and Behavioral Sciences, 191, 1479–1482.

Ouahbi, I., Darhmaoui, H., Kaddari, F., Elachqar, A., & Lahmine, S. (2015b). Vers un enseignement de la programmation compatible avec la culture vidéoludique des élèves au Maroc. Actes de la 7ème Conférence sur les Environnements Informatiques d'Apprentissage Humain EIAH'15, Agadir, Maroc

Prensky, M. (2007). Digital Game-based Learning (Paragon House Ed.). Paragon House Publishers.

Romero, M., & Barma, S. (2015). Teaching pre-service teachers to integrate Serious Games in the primary education curriculum. International Journal of Serious Games, 2(1).

Sawyer, B., & Smith, P. (2008). Serious Games Taxonomy. Presented at the Serious Games Summit 2008, San Fransisco, USA.

Shaffer, D. W. (2006). How computer games help children learn. Palgrave Macmillan.

Instrumentations pour l'étude des consonnes géminées du tarifit

Fayssal Bouarourou[1] Béatrice Vaxelaire[1] Yves Laprie[2]
Rachid Ridouane[3] Rudolph Sock[4-1]
[1] U.R. 1339 LILPA/ Equipe de Recherche Parole et Cognition & Institut de Phonétique de Strasbourg (IPS), Université de Strasbourg (Uds), 22, rue Descartes 67084 Strasbourg, France.
[2] LORIA/CNRS, Nancy, France.
[3] Laboratoire de Phonétique et Phonologie (CNRS-Sorbonne nouvelle), Paris, France.
[4] Language Information and Communication Laboratory- LICOLAB -Université Pavla JozefaSafarika, Faculté des Lettres Košice, Slovaquie.
fayssalbouarourou@gmail.com

Résumé

L'objectif principal de cette étude est de présenter les méthodes instrumentales acoustique et articulatoire que nous utilisons dans le domaine de la phonologie de laboratoire pour l'étude de la gémination en tarifit. Cette étude repose sur des données acoustiques pour six locuteurs natifs et sur des données cinéradiographiques pour deux locuteurs natifs. Elle présente les résultats de consonnes simples et géminées, sourdes et sonores, en position initiale, intervocalique et finale, en vitesses d'élocution normale et rapide. La vitesse d'élocution est variée afin d'évaluer la robustesse du contraste phonologique. Une attention particulière est accordée à la synchronisation des gestes de la langue dans la production du contraste phonologique. Le corpus de la présente investigation, qui fait suite à une précédente étude (Bouarourou, 2014), nous permet de comparer les patterns spatiotemporels articulatoires et acoustiques des géminées à ceux des géminées hétéromorphémiqueshomorganiques du tarifit. C'est seulement en procédant à une analyse comparative de telles données que nous pourrons savoir si, d'un côté nous avons un contrôle spatiotemporel correspondant à «doing the samething once, but for a longer period of time » (faire la même chose une fois, mais sur un intervalle plus long), ou de l'autre côté, un contrôle spatiotemporel renvoyant à «doingtwothings at once » (« faire deux choses à la fois »). Toujours par rapport au corpus, nous pourrons savoir si les séquences cibles en position intervocalique ne posent aucun problème, celles en position initiale et finale n'étant pas en positions absolues. Nous pourrons donc tester, au sens strict du terme, l'effet de ces positions sur le trait phonologique. Au final, cette étude nous permettra de savoir si les caractéristiques phonétiques spatio-temporelles sous-jacentes à la production des géminées en tarifit peuvent être formalisées, d'un point de vue phonologique, par une représentation structurelle qui considérerait la gemination comme le timing de deux gestes consécutifs associés à un segment temporelle simultané.

1. Introduction

L'objectif de notre travail est d'étudier le phénomène phonologique de la gémination en berbère du tarifit parlé au Maroc, dans la province de Nador. L'étude repose sur des investigations acoustiques et articulatoires, menées principalement dans le domaine temporel. Il s'agit essentiellement d'identifier des indices acoustiques et articulatoires potentiels qui pourraient sous-tendre le trait phonologique de la gémination.

Quelques hypothèses seront formulées, aussi bien dans le domaine acoustique que dans le domaine articulatoire. De manière générale, nous pensons que l'analyse du substrat physique, articulatoire et acoustique, devrait nous permettre de mettre au jour des indices articulatoires et acoustiques qui sont sous-jacents au trait phonologique. La gémination étant principalement un phénomène temporel, ces indices devraient être plus remarquables au niveau du contrôle temporel des gestes articulatoires et de leurs conséquences acoustiques.

2. Etude acoustique

Sur le plan acoustico-perceptif, il est important de citer les travaux précédents (par ex. Ridouane, 2003 ; Abramson, 1987, 1998 ; Lahiri & Hankamer, 1988 ; Arvaniti & Tserdanelis, 2000, 2001 ; Kawahara, 2012), menés sur différenteslanguespossèdant la distinction simple / géminée, afin de dégagerles indices acoustiquesresponsables de cette opposition. Plusieurs tests ont été effectués sur plusieurs paramètres comme la durée des voyelles adjacentes, la durée de l'occlusion, la durée du VOT (Voice Onset Time ou délai d'établissement du voisement), la durée des fricatives, l'amplitude du burst (explosion-friction) et la fréquence fondamentale de la voyelle suivante. Ces auteurs ont dégagé plusieurs indices qui permettent de distinguer les simples des géminées. La plupart des travaux cités trouvent que la durée de la tenue consonantique est l'indice principal pour distinguer les deux catégories. D'autres indices trouvés sur certaines langues restent secondaires.

Notre analyse du signal acoustique dans notre étude sur le tarifit repose sur une approche articulatori-acoustique. Celle-ci nous permet de déceler, sur le signal de parole, des événements acoustiques interprétables en terme articulatoire (Abry *et al.*, 1985). À partir de cette approche événementielle, nous déterminons des intervalles spécifiques, révélateurs de l'organisation temporelle ou timing du signal de parole, et pouvant nous permettre de retrouver dans ce substrat physique des indices qui pourraient sous-tendre l'opposition phonologique de la gémination.

Dans cette étude, nous avons pu définir les paramètres acoustiques qui pourraient permettre de distinguer les consonnes simples des géminées. Nous avons enregistré six locuteurs. Chaque locuteur a prononcé, dans un ordre aléatoire, un corpus de 27 paires minimales, répété dix fois, d'abord en vitesse d'élocution normale, puis en vitesse d'élocution rapide. Les mots ont été insérés dans une phrase porteuse : "Ini Ê iZ umar" qui signifie "dis Ê une fois", cela pour contrôler, autant que possible, le contexte segmental et prosodique et donner un certain sens aux phrases. Le corpus comporte cinq paires d'occlusives : trois sourdes et deux sonores. Il comporte également quatre paires de constrictives : deux sourdes et deux sonores. Ces paires ont été introduites en trois positions de mot : initiale (non absolue), intervocalique et finale (non absolue). Les sujets sont deux locutrices agées de 25 et 26 ans et quatre locuteurs

agés de 27 à 35 ans, au moment de l'enregistrement. Les enregistrements ont été effectués à l'Institut de Phonétique de Strasbourg dans la chambre sourde, pour deux locuteurs (F et Kh), et au Maroc pour les quatre autres locuteurs (K, Y, H et S) dans des endroits silencieux. À l'aide du logiciel PRAAT, nous avons segmenté le signal acoustique et mesuré les paramètres (intersegmentaux et intrasegmentaux) suivants : a) la durée de la voyelle précédant la consonne cible, simple ou géminée ; b) la durée du VTT (Voice Termination Time ou délai d'arrêt du voisement) ; c) la durée de l'occlusion ou du silence acoustique ; d) la durée du VOT ; e) la durée de la friction ; f) la durée de la voyelle qui suit la consonne cible. Toutes les mesures ont été soumises à des analyses statistiques (ANOVA).

Dans cette investigation, nous allons aussi analyser quelques paramètres temporels des occlusives voisées et non voisées et des constrictives voisées et non voisées, dans trois positions de mot : initiale non absolue, intervocalique et finale non absolue. Les mesures de durées absolues ont été effectuées, d'une part au niveau intersegmental, à savoir la durée de la tenue consonantique et la durée des voyelles adjacentes et, d'autre part, au niveau intrasegmental, à savoir la durée du VTT, du silence acoustique et du VOT (Klatt, 1975) pour les occlusives sourdes (voir Figure 1). En ce qui concerne les occlusives sonores, outre les durées intersegmentales de la tenue consonantique et des voyelles adjacentes, nous avons mesuré la durée de l'occlusion et celle du VOT (voir Figure 2). L'objectif de cette expérience est de voir si les paramètres temporels mentionnés ci-dessus jouent un rôle significatif, au niveau du timing, dans l'opposition phonologique entre les consonnes simples et les géminées du tarifit. Il s'agit aussi de tester la robustesse de ces paramètres dans le maintien éventuel de l'opposition phonologique, en variant la vitesse d'élocution (Bouarourou, 2010).

2.1. Hypothèses

- *Hypothèse n° 1* : La gémination étant un fait phonologique temporel de quantité consonantique, il est tout-à-fait cohérent de s'attendre à des différences de durées consonantiques, plus longues pour les géminées que pour les simples.
- *Hypothèse n° 2* : Des faits phonologiques reposant habituellement sur des dimensions pluri-indicielles, nous pensons que la ou les voyelles adjacentes pourraient contribuer à renforcer le trait phonologique de la gémination.
- Le silence acoustique pour les sourdes (*Hypothèse n° 3*) et l'occlusion consonantique pour les sonores (*Hypothèse n° 4*), deux intervalles intrasegmentaux consonantiques, devraient sous-tendre l'opposition phonologique de la gémination, en tant qu'indices de renfort du trait potentiellement principal de la tenue consonantique.
- *Hypothèse n° 5* : Nous pensons, comme précisé auparavant, que l'opposition entre consonnes simples et géminées, qui pourrait reposer aussi sur des différences de pression intra-orale durant la tenue consonantique, et sur des différences de changement de la pression d'air au relâchement, aurait une incidence sur la durée du VOT ; la durée du VOT des géminées devrait, en conséquence, être plus longue que celle des simples.
- *Hypothèse n° 6* : Il est probable que le VTT des simples soit plus long que celui des géminées, étant donné que plus la tenue consonantique est brève, plus la proportion du délai d'arrêt du voisement dans cette tenue risque d'être élevée.

- *Hypothèse n° 7* : Le signal de parole étant intrinsèquement élastique, tous les segments devraient subir une compression avec l'augmentation de la vitesse d'élocution, ce qui ne devrait pas empêcher les catégories oppositives de rester distinctes, afin de préserver l'opposition phonologique de la gémination.

2.2. Données acoustiques Absolues

Au niveau acoustique, les données absolues reposent sur les mesures suivantes : a) la durée des consonnes simples et géminées dans trois positions : intervocalique, initiale non absolue et finale non absolue, produites à deux vitesses d'élocution, normale et rapide ; b) les durées intersegmentales pour les consonnes simples et géminées dans les trois positions citées ci-dessus ; c) les durées intrasegmentales des occlusives sourdes et sonores en vitesse d'élocution normale et rapide (ces durées intrasegementales ont été prélevées pour les sourdes en position intervocalique, en initiale et finale non absolue, et uniquement en position intervocalique et initiale non absolue pour les sonores).

Les résultats dans le domaine intersegmental (Figure 1) montrent, pour toutes les consonnes produites en vitesse d'élocution normale ou rapide, que la durée de la tenue consonantique est significativement plus longue pour les géminées que pour les simples (pour chaque consonne analysée *p<0,0001*) (conformément à l'*Hypothèse n° 1*).

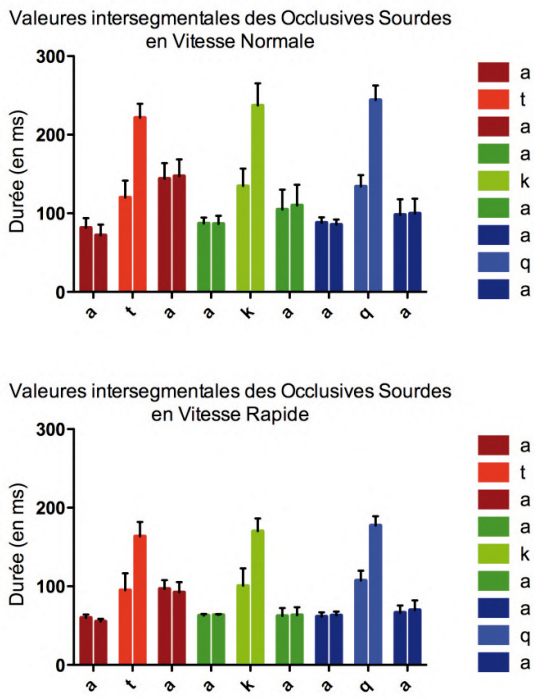


Figure 1 : Valeurs intersegmentales des occlusives sourdes (VCV)envitesse d'élocution normale à gauche et en vitesse d'élocution rapide à droite.

Ce résultat corrobore ceux attestés dans la littérature (Ridouane, 2003, 2007 ; Abramson, 1987, 1998 ; Lahiri & Hankamer, 1988). La durée des voyelles adjacentes ne contribue pas à l'opposition phonologique de la gémination (l'Hypothèse n° 2 n'est pas vérifiée). Nos données pour le tarifit ne correspondent pas à celles de McKay (1980) sur le rembarrnga (langue nord-australienne), qui ont démontré que les voyelles sont plus courtes devant les géminées que devant les simples. Elles ne correspondent pas non plus à celles deCampbell (1999), Fukui (1987), Han (1994), Hirata (2007), Hirose & Ashby (2007), Idemaru & Guion (2008), Kawahara (2006), Ofuka (2003), Port *et al.* (1987) et Takeyasu (2012), qui ont montré que les voyelles sont plus longues devant les géminées que devant les simples (cité par Kawahara, 2012). En revanche, elles confirment celles attestées par Ridouane (2003; 2007), Abramson (1987; 1998) et Lahiri & Hankamer (1988).

Dans le domaine intrasegmental (Figure 2), l'analyse a permisd'observer que la durée du silence acoustique des géminées pour les occlusivessourdes est significativement plus long que celle de leurs homologues simples (pour chaque consonne analyséep<0,0001) (l'Hypothèse n° 3 est vérifiée). Pour les occlusives sonores simples et géminées, c'est la durée de l'occlusion qui permet de distinguer les deux catégories (pour chaque consonne analysée p<0,0001) (ce qui est conforme à l'Hypothèse n° 4). Le VOT n'intervient pas dans la séparation des catégories phonologiques (ce qui va à l'encontre de l'Hypothèse n° 5). Il en va de même du VTT qui ne permet pas de séparer les occlusives sourdes simples des géminées (Hypothèse n° 6 n'est pas vérifiée). La compression des segments, provoquée par l'augmentation de la vitesse d'élocution, ne nuit pas au maintien du contraste phonologique (l'Hypothèse n° 7 est confirmée).

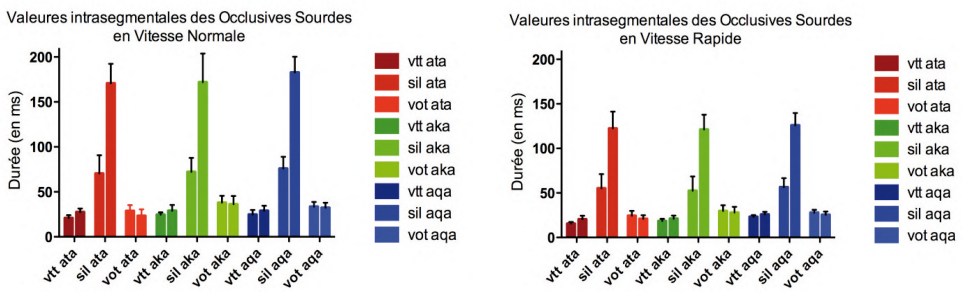


Figure 2 : Valeurs intrasegmentales des occlusives sourdes (VCV) en vitesse d'élocution normale à gauche et en vitesse d'élocution rapide à droite.

3. Etude articulatoire

Sur le plan articulatoire et physiologique, plusieurs études ont été menés sur différentes langues qui possèdent la distinction simple / géminée (par ex. Vaxelaire, 1995 ; Ridouane, 2003 ; 2007 ; Farnetani, 1990 ; Byrd, 1995 ; Kraehenmann & Lahiri, 2008 ; Löfqvist, 2005; 2006 ; 2007 ; 2009), afin de dégager les indices articulatoires responsables de cette opposition. Différentes méthodes ont été testées sur le phénomène de la gémination aux niveaux glottique et supraglottique. La méthode fibroscopique permet de tester la durée et le degré d'ouverture

glottale pendant la tenue des consonnes simples et des consonnes géminées. Quant à la méthode électromyographique, elle s'appuie sur le nombre de sommets des courbes pour distinguer les simples des géminées et pour vérifier si les géminées impliquent une ré-articulation ou non. En ce qui concerne la méthode électropalatographique, elle est utilisée pour déterminer la durée et l'étendue du contact entre la langue et le palais. Enfin, la méthode cinématique vise à observer les différents contacts et les différents mouvements des articulateurs, afin de comparer les consonnes simples aux consonnes géminées. Les principaux enseignements à tirer de telles investigations sont les suivants : a) la durée acoustique de la tenue consonantique est plus longue pour les géminées ; b) des occlusions (contacts ou étendues de contact) sont plus longues pour les géminées ; c) la forte augmentation de la pression intraorale ; un mouvement de la langue de la première à la deuxième voyelle ayant une durée plus longue pour les consonnes longues que pour les brèves. La radiocinématographie a contribué à une avancée considérable pour l'étude des gestes des articulateurs, et notamment ceux situés à l'intérieur du conduit vocal. Elle apporte une contribution décisive au développement de théories sur la production de la parole, et aux premiers essais de synthèse de type articulatoire.

Notre étude articulatoire est basée sur une étude cinéradiographique pour l'étude du phénomène de gémination du tarifit. Dans ce travail, nous avons étudié deux films des deux locuteurs KH et F (un par locuteur) que nous avons extraits de la base de données de l'Institut de Phonétique de Strasbourg pour la présente étude (les films sont au format DICOM (25 Hz)). Une post-synchronisation des images et du son est nécessaire avant de pouvoir exploiter les données. Les radiofilms fournissent une information dans le plan sagittal de la position et des déplacements d'un certain nombre d'articulateurs tels que les lèvres, la mâchoire, la langue, le voile du palais, le pharynx, l'os hyoïde et le larynx (voir Figure 3). Le logiciel X-Articulator nous a permis d'obtenir le contour des différents articulateurs de manière automatique ou semi-automatique sur les films de cette investigation. Pour notre étude sur la gémination, nous avons retenu les paramètres articulatoires suivants : a) l'ouverture de la constriction ; b) la constriction pharyngale ; c) l'aperture labiale ; d) la position du larynx ; e) la position de l'os hyoïde ; f) l'étendue de contact.

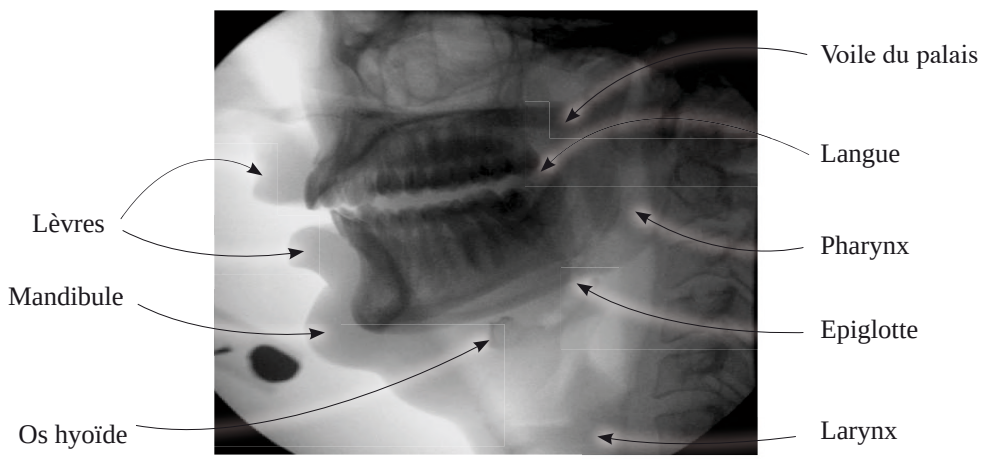


Figure 3 : Radiographie d'une vue sagittale du conduit vocal

3.1. Hypothèses

- *Hypothèse n° 1* : Les différences de duréeentre les consonnes géminées et les consonnes simples observées sur le plan acoustique devraient aussi être visibles prioritairement au niveau du contrôle temporel des paramètres articulatoires.
- *Hypothèse n° 2* : Ces différences de durée pourraient être accompagnées de différences au niveau des déplacements d'articulateurs critiques dans la dimension spatiale.
- *Hypothèse n° 3* : Desdifférences spatiotemporelles entre nos deux locuteurs pourraient être observées dans cette investigation, si l'on prend en compte les facteurs liés généralement aux spécificités du locuteur.
- *Hypothèse n° 4* : Nous examinerons les phénomènes d'anticipation. Supposant que pour la réalisation de gestes potentiellement plus longs pour les géminées, il soit nécessaire d'initier ceux-ci plus tôt. En corolaire, il est probable que le démarrage des gestes vocaliques post-consonantiques soit :
  (a) *retardé* en contexte géminé, le temps de réaliser correctement une striction (contact ou constriction) longue, mais
  (b) *anticipé* en contexte d'une consonne simple, où la striction dure relativement moins longtemps.
- *Hypothèse n° 5* : Au niveau glottique, la différence entre simples et géminées serait la même que celle qui oppose la production de consonnes implosives et éjectives. L'ensemble *larynx-os hyoïde* devrait être plus élevé pour les géminées, comme pour les éjectives, à cause d'un taux de la pression intra-orale supérieur pour les géminées.

3.2. Données articulatoires

L'analyse des données de nos deux locuteurs, Kh et F, pour les séquences VCV, où la consonne est une occlusive ou une constrictive sourde ou sonore, révèle que nos paramètres articulatoires, à savoir l'ouverture de la constriction, la constriction pharyngale et l'aperture labiale sont de bons candidats pour l'étude de l'organisation spatiotemporelle de la production de la consonne simple et de son homologue géminée, aussi bien dans le contexte alvéolaire que vélaire et uvulaire (pour les sourdes). La trajectoire de ces trois paramètres, dans les configurations consonantiques cibles, est structurellement similaire entre la simple et la géminée.

Au niveau spatial, si la réduction de l'ouverture de la constriction pour réaliser les contacts alvéolaires et vélaires est toujours positivement « corrélée » avec la réduction de l'aperture labiale, ces deux derniers paramètres sont inversement « corrélés » avec la constriction pharyngale ; celle-ci augmente lorsque les deux autres paramètres diminuent. Pour ce qui concerne la consonne uvulaire /q/, en revanche, la constriction pharyngale diminue en même temps que l'ouverture de la constriction, alors que l'aperture labiale, elle, reste relativement stable. Nous posons, au vu de ces résultats, que l'augmentation de la taille de la constriction au niveau de la cavité pharyngale, aussi bien lors de la production de la consonne simple que de son homologue géminée, n'est pas nécessairement attribuable au geste de protrusion de

la langue, vers la partie frontale de la cavité buccale, dans la réalisation du contact apical. En effet, cette augmentation de la taille de la cavité pharyngale est aussi clairement visible en contexte vélaire, un contexte qui ne favorise pas particulièrement une avancée de la masse linguale. On peut donc, dans ce cas, supposer l'existence d'un geste typiquement consonantique (*Hypothèse n° 2*), avec la taille d'une constriction pharyngale spécifique, et un autre geste typiquement vocalique, ayant une constriction pharyngale réduite, intrinsèque à la voyelle radico-pharyngale /a/. Reste maintenant à vérifier cette supposition à partir d'observations en contextes vocaliques d'une voyelle antérieure, telle que le /i/, par exemple.

Par rapport au timing des trois premiers paramètres articulatoires (l'ouverture de la constriction, la constriction pharyngale et l'aperture labiale) illustré dans les Figures 4 et 5 ci-dessus, il est évident que le contact alvéolaire, vélaire et uvulaire, plus long pour la géminée, est le paramètre préférentiel de la distinction phonologique simple vs.géminée (*Hypothèse n° 1*). Dans le cas des consonnes alvéolaires et vélaires, ce contact est accompagné de l'augmentation de la constriction pharyngale et de la réduction de l'aperture labiale, qui durent plus longtemps pour la géminée par rapport à la simple, pendant l'occlusion. Le contact uvulaire, lui, est accompagné d'une réduction de la constriction pharyngale, qui a une durée plus longue pour les géminées, comparées aux simples. L'aperture labiale, pour l'uvulaire, reste relativement stable tout au long de la production de la séquence VCV.

Toujours au niveau du timing articulatoire, nous avons constaté que les gestes anticipatoires, pour produire les consonnes cibles, simples ou géminées, sont initiés plus précocement durant la production de la voyelle précédente. Cette stratégie anticipatoire est encore plus précoce pour les consonnes géminées, comparées aux consonnes simples en contextes alvéolaire et vélaire. Nous avons constaté aussi, dans ces contextes, la mise en place d'une stratégie de coarticulation rétentrice au service de la réalisation de la gémination, afin de marquer le contraste temporel entre les deux termes de l'opposition phonologique. Cette dernière conduite motrice n'est pas généralisable en contexte uvulaire (*Hypothèse n° 4*).

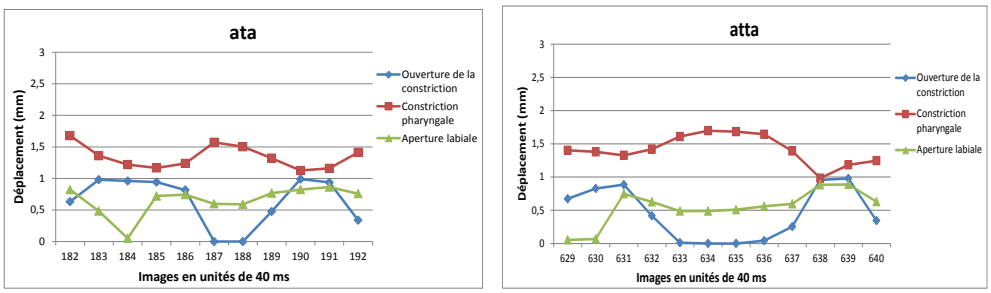


Figure 4 : Trajectoires de trois paramètres articulatoires, à savoir l'ouverture de la constriction, la constriction pharyngale et l'aperture labiale, pour la séquence /ata/ (à gauche) et /atta/ (à droite) ; locuteur F

Par rapport aux constrictives sourdes et sonores, l'examen des trajectoires de nos trois paramètres articulatoires dans les contextes révèle des comportements clairs et généralisables.

Au niveau spatial, l'agrandissement du pharynx lors de la réalisation de la consonne, observé

pour les occlusives, se confirme ici pour les constrictives, même si le constat est moins prononcé et moins systématique pour les consonnes simples (*Hypothèse n° 2*). L'aperture labiale diminue ou reste stable en même temps que la réduction de la constriction alvéolaire.

Au niveau du timing des gestes, la constriction maximale dure plus longtemps pour les géminées que pour les simples, et cela de façon remarquable (*Hypothèse n° 1*). La diminution de l'aperture labiale durant le maintien de la constriction alvéolaire maximale dure, elle aussi, plus longtemps pour les géminées que pour les simples.

Les gestes anticipatoires sont plus nets lors des transitions de voyelles à consonnes que des transitions des consonnes cibles aux voyelles suivantes. C'est grâce à ces stratégies de persévération des gestes associés aux configurations fermantes du conduit vocal pour les géminées que les différences temporelles entre consonnes simples et consonnes géminées sont maximisées (*Hypothèse n° 4*).

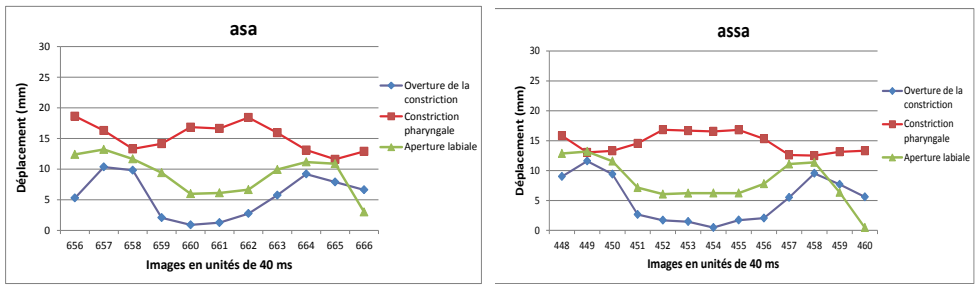


Figure 5 : Trajectoires de trois paramètres articulatoires, à savoir l'ouverture de la constriction, la constriction pharyngale et l'aperture labiale, pour la séquence /asa/ (à gauche) et /assa/ (à droite) ; locuteur Kh

En ce qui concerne le niveau laryngal, nous avons remarquéune similarité entre les trajectoires du larynx et de l'os hyoïde à travers nos différentes séquences, ainsi qu'une ressemblance structurelle des trajectoires, quels que soient les contextes consonantiques, les positions des séquences cibles dans les mots et les locuteurs. Ce couplage étroit observé entre le couple larynx-os hyoïde se confirme même si la « corrélation » entre les trajectoires de ces deux structures est obtenue à partir de leur déplacement vertical respectif. Rappelons que Vaxelaire et Sock (1997) obtiennent, eux, des corrélations relativement élevées entre le déplacement vertical du larynx et la composante horizontale du mouvement diagonal de l'os hyoïde.

Nous avons trouvé que le contrôle spatiotemporel du geste du larynx et de celui de l'os hyoïde était comparable entre les consonnes simples et les consonnes géminées (*Hypothèse n° 5* n'est pas confirmée). Le mouvement de l'ensemble larynx-os hyoïde est inversement corrélé avec le déplacement vertical de la masse de la langue et, en conséquence, avec l'aperture vocalique. Ainsi,nos données rejettent ellesaussi « l'hypothèse tongue-pull » (*tongue-pull hypothesis*). L'analyse du déplacement vertical de l'ensemble larynx-os hyoïde ne permet pas de déceler un comportement qui serait à lier avec des oppositions consonantiques phonologiques sourdes vs.sonores. Cette affirmation devrait toutefois être nuancée étant donné que le contexte adjacent n'estpas comparable.

De manière générale, nous ne trouvons pas de spécificités spatiotemporelles par rapport aux locuteurs (contrairement à l'*Hypothèse n° 3*). L'analyse des vues de profil indique les résultats suivants sur l'étendue de contact : a) l'étendue de contact des occlusives est systématiquement plus importante pour les géminées que pour les simples ; b) l'étendue de contact augmente de la consonne alvéolaire, à la vélaire (réalisée plutôt palatale ici), puis à l'uvulaire (à quelques rares exceptions près). Par rapport à la durée de la tenue articulatoire (en ms), celle des occlusives et des constrictives géminées est plus longue que celle de leurs homologues simples.

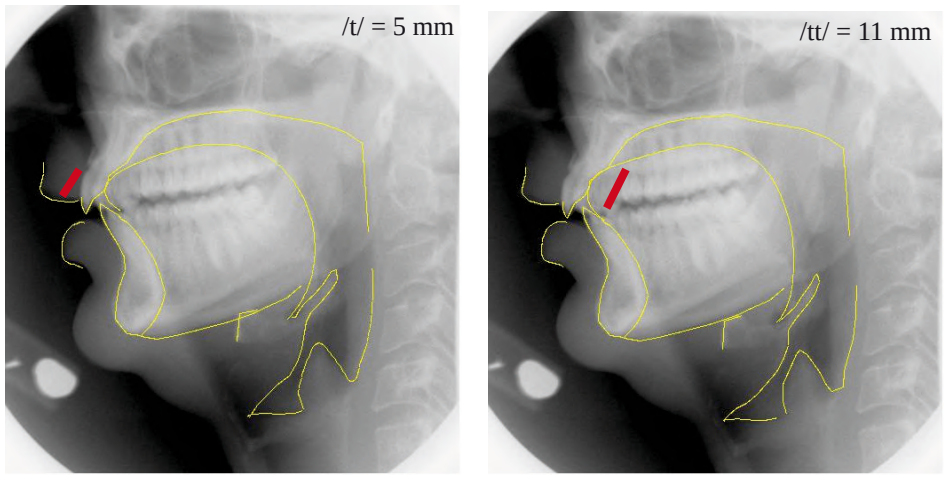


Figure 6 : l'étendu de contact de /t/ = 5 mm (à gauche) et /tt/ = 11 mm (à droite) ; locuteur F

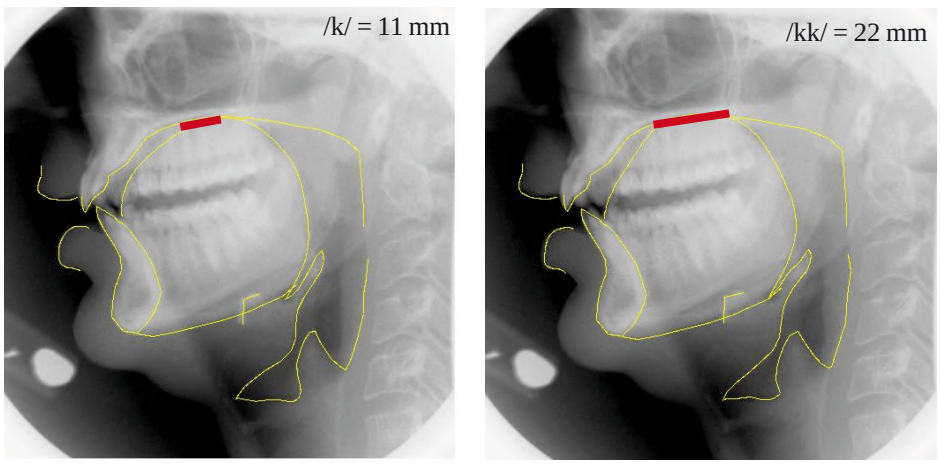


Figure 7 : l'étendu de contact de /k/ = 11 mm (à gauche) et /kk/ = 22 mm (à droite) ; locuteur F

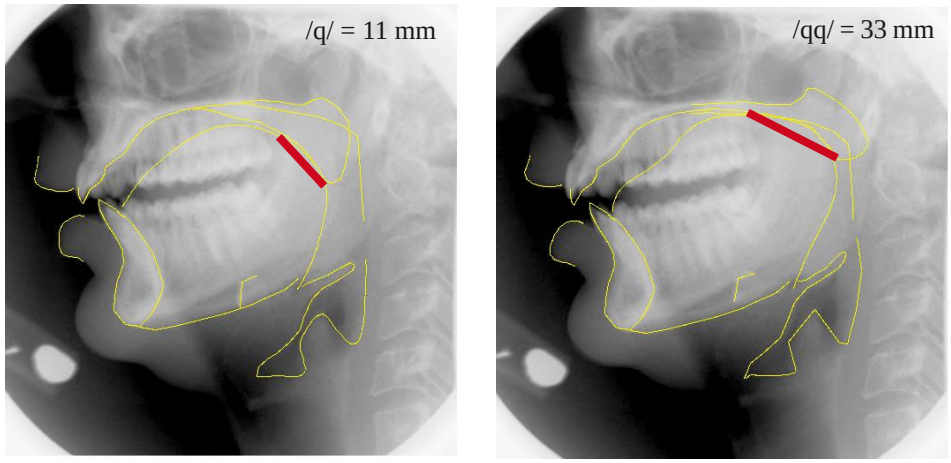


Figure 8 : l'étendu de contact de /q/ = 11 mm (à gauche) et /qq/ = 33 mm (à droite) ; locuteur F

Au niveau des relations entre paramètres articulatoires, les données montrent une sorte d'assignation de valeurs de base : à tenue articulatoire longue, une étendue de contact large, et vice versa. Par rapport aux relations articulatori-acoustiques, on remarque que la combinaison de l'étendue de contact et de la durée acoustique des consonnes permet une nette distinction des termes de l'opposition phonologique. L'examen des données révèle que lorsque deux paramètres temporels sont combinés, à savoir la tenue articulatoire et la durée acoustique (Figure 9), ils possèdent une puissance notoire pour distinguer nettement les deux catégories phonologiques, aussi bien pour les consonnes non voisées que voisées, et autant pour les occlusives que pour les constrictives. Le phénomène phonologique de la gémination repose décidemment et prioritairement sur le substrat physique temporel.

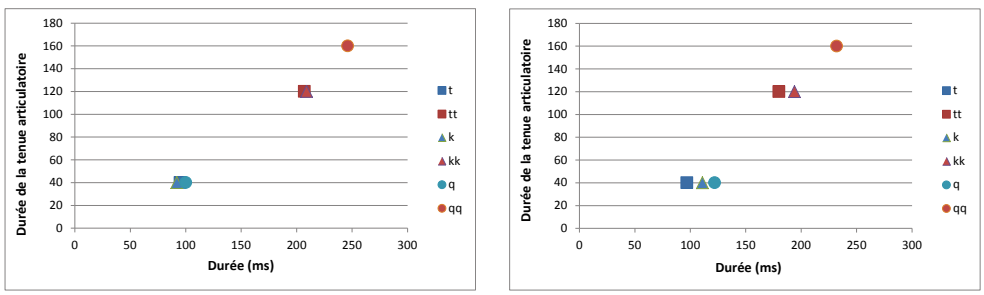


Figure 9 : Relations entre la tenue articulatoire et la durée acoustique des occlusives sourdes; locuteur F (à gauche) et locuteur KH (à droite)

4. Conclusion

Un certain nombre d'hypothèses a été formulé, aussi bien dans le domaine acoustique que dans le domaine articulatoire. Il s'agissait essentiellement d'identifier des indices acoustiques et articulatoires potentiels qui pourraient sous-tendre le trait phonologique de la gémination.

Sur le plan acoustique, nos valeurs intersegmentales nous a montré que : 1) La tenue consonantique a été systématiquement plus longues pour les consonnes géminées que pour les simples, tous locuteurs, contextes consonantiques et conditions de vitesse d'élocution confondus ; 2) Le contexte des voyelles adjacentes ne semble pas contribuer à la distinction entre simples et géminées sur la dimension temporelle.

D'après nos valeurs intrasegmentales, nous avons pu constater que la durée du silence acoustique pour les sourdes et la durée d'occlusion consonantique pour les sonores, deux intervalles intrasegmentaux consonantiques, sont des indices principaux de l'opposition phonologique de la gémination. Ces valeurs nous ont montré aussi que ni le VOT, ni le VTT ne contribuent à cette opposition simple/géminée et que la compression provoquée par l'augmentation de la vitesse d'élocution concerne tous les segments. Malgré cette compression, nos catégories restent distinctes.

Sur le plan articulatoire, notre étude cinéradiographique sur le phénomène phonologique de la gémination du tarifit nous a montré que : 1) les différences de durée entre les consonnes géminées et les consonnes simples, observées sur le plan acoustique, ont été visibles au niveau du contrôle temporel des paramètres articulatoires retenus pour cette investigation, puisque les occlusions articulatoires et la tenue des constrictions articulatoires ont toutes été systématiquement plus longues pour les géminées que pour les simples ; 2) des caractéristiques articulatoires spatiales, attribuables aux configurations consonantiques fermantes des consonnes cibles ont certes été décelées mais seule l'étendue de contact permet de manière robuste de séparer les simples des longues ; 3) les deux locuteurs ont comparativement les mêmes stratégies articulatoires, en termes de contrôle spatiotemporel, pour réaliser les différences entre consonnes simples et consonnes géminée ; 4) il est vrai que des stratégies anticipatoires, tout comme des faits liés à la rétention spatiotemporelle des gestes, servent à faire émerger correctement les géminées ; 5) la position du couple larynx-os hyoïde reste comparable pour les deux termes de l'opposition phonologique. Cela semble indiquer que des différences éventuelles de pression intra-orale n'ont pas eu d'incidence sur la position de ce couple et, en conséquence, sur les valeurs du VOT.

Au total, nos données et nos observations nous amènent à insister sur la mise au jour de régularités acoustiques et articulatoires dans la réalisation du trait phonologique de la gémination. Si l'on ne peut pas, selon nos résultats, parler de présence d'invariances absolues ou relatives qui sous-tendraient la gémination, nous pensons qu'il convient de prendre en compte les indices acoustiques robustes - que nous avons signalés - présents dans le signal de parole (Stevens, 1972), ainsi que les indices articulatoires relatifs au niveau du contrôle spatiotemporel des gestes articulatoires (Liberman & Mattingly, 1985; Fujimura,1991; Browman & Goldstein, 1989 ; Saltzman *et al.*, 1987). On gardera toutefois à l'esprit la nature écologique du système de production / perception de la parole qui devrait forcément intégrer les phénomènes de variabilités (Lindblom, 1987) pour être mieux appréhendé.

C'est dans cette perspective de relations articulatori-acoustiques que nous proposons de traiter le phénomène phonologique de la gémination en tarifit comme suit :

1) Les consonnes simples du tarifit correspondraient à :

   a) un seul geste ayant son organisation spatiotemporel spécifique : une occlusion / constriction supraglottique d'une certaine durée, accompagnée d'une étendue de contact d'une largeur déterminée ;

   b) une tenue consonantique acoustique, renforcée par un silence acoustique ou par une occlusion acoustique de durée spécifique.

2) Les géminées correspondraient à :

   a) un geste unique comparable à celui de son homologue simple, mais ayant une occlusion / constriction supraglottique d'une durée supérieure à celle de la simple, accompagnée d'une étendue de contact d'une largeur également supérieure à celle de la simple ;

   b) une tenue consonantique acoustique, renforcée par un silence acoustique ou par une occlusion acoustique de durée supérieure à celle de son homologue simple.

Ainsi, les géminées du tarifit seraient calquées sur leurs homologues simples, le locuteur ayant appris à mettre en place des réorganisations spatiotemporelles articulatoires et acoustiques adéquates pour passer d'une entité à l'autre. En effet, selon nos données articulatoires et acoustiques, le locuteur *ferait la même chose une fois, mais pour un lapse de temps plus long (« doing the samething once, for a longer period of time* »), puisque :

1) les patterns spatiotemporels articulatoires sont structurellement comparables entre simples et géminées ;
2) seul un contact unique et prolongé a été observé pour les géminées ;
3) la nature de ce contact est qualitativement mais pas quantitativement similaire à celui des simples ;
4) le signal acoustique, en conséquence, ne révèle aucune rupture ni dans la phase silencieuse de la tenue consonantique des sourdes, ni dans celle de l'occlusion consonantique des sonores.
5) La gémination en tarifit serait un phénomène, entre autres, de réaménagement (rescaling) temporel des phasages et de réorganisation d'un paramètre spatial critique.

Remerciements

Ce travail a été financé par : a) un programme de la Maison Interuniversitaire des Sciences de l'Homme Alsace (MISHA), 2013-2016, « Percevoir la parole : une histoire sensori-motrice d'événements audibles et visibles » ; b) un projet Attractivité IdEx (Initiative d'Excellence) de l'Université de Strasbourg, Unistra, 2014-2016, « Plateforme Unistra de Linguistique et Phonétique Cliniques – PULPC », attribués à l'Institut de Phonétique de Strasbourg / E.A. 1339 Linguistique, Langues et Parole - U.R. LiLPa, Equipe de Recherche – E.R. Parole et Cognition.

Références

Abramson, A. S. (1987). Word-initial consonant length in Pattani Malay. *In Proceedings of the 11th International Congress of Phonetic Sciences* (Vol. 6, pp. 143–147). Tallinn.

Abramson, A. S. (1998). The complex acoustic output of a single articulatory gesture: Pattani Malay word-initial consonant length. In U. Warotamasikkhadit& T. Panakul (Eds.), *Papers from the Fourth Annual Meeting of the Southeast Asian Linguistics Society* 1994 (pp. 1–20). Tempe, Arizona.

Abry, C., Autesserre, D., Barrera, C., Benoit, C., Boë, L. J., Caelen, J., … Vigouroux, N. (1985). Propositions pour la segmentation et l'étiquetage des sons du français. In *14e JEP du GCP du GALF* (pp. 156–163).

Arvaniti, A., &Tserdanelis, G. (2000). On the phonetics of geminates: evidence from Cypriot Greek. *In Proceedings of the 6th International conference on Spoken Language Processing* (Vol. 2, pp. 559–562). Beijing.

Arvaniti, A., &Tserdanelis, G. (2001). The acoustic characteristics of geminate consonants in Cypriot Greek. *In Proceedings of the 4th International Conference on Greek Linguistics* (pp. 29–36). Thessaloniki

Bouarourou, F., Vaxelaire, B., Ridouane, R., Hirsch, F., &Sock, R. (2010). La résistivité de la gémination en tarifit. In Actes des 28e Journées d'Étude sur la Parole (pp. 341–345). Mons.

Bouarourou, F. (2014). *La gémination en tarifit : considérations phonologiques, études acoustique* et articulatoire (Thèse de doctorat). Université de Strasbourg, France.

Browman, C. P., & Goldstein, L. (1989). Articulatory gestures as phonological units. *Haskins Laboratories Status Report on Speech Research,* 99/100, 69–101.

Byrd, D. (1995). Articulatory characteristics of single and blended lingual gestures. In *Proceedings of the 13th International Congress of Phonetic Sciences 2* (pp. 438–441). Stockholm.

Campbell, N. (1999). A study of Japanese speech timing from the syllable perspective. Onsei Kenkyuu. [*Journal of the Phonetic Society of Japan*], 3(2), 29–39.

Farnetani, E. (1990). V-C-V Lingual Coarticulation and Its Spatiotemporal Domain. In W. J. Hardcastle & A. Marchal (Eds.), *Speech Production and Speech Modelling* (pp. 93–130). Springer Netherlands.

Fujimura, O. (1991). Beyond the segment. In I. G. Mattingly & M. Studdert-Kennedy (Eds.), *Modularity and the Motor Theory of Speech Perception* (pp. 25–31). Hillsdale, New Jersey: Lawrence Erlbaum.

Fukui, S. (1987). Nihongo heisaon-no encho/tanshuku-niyoru sokuon/hisokuon toshite-no chooshu [perception for the Japanese stop consonants with reduced and extended durations]. *Onsei Gakkai Kaihou [Phonetic Society Reports*], 59, 9–12.

Han, M. S. (1994). Acoustic manifestations of mora timing in Japanese. *The Journal of the Acoustical Society of America, 96*(1), 73–82.

Hirata, Y. (2007). Durational variability and invariance in japanese stop quantity distinction: Roles of adjacent vowels. Onsei Kenkyuu [*Journal of the Phonetic Society of Japan*], 11(1), 9–22.

Hirose, A., & Ashby, M. (2007). An acoustic study of devoicing of the geminate obstruents in Japanese. *In Proceedings of the XVIth International Congress of Phonetic Sciences* (pp. 909–912). Saarbrücken.

Idemaru, K., & Guion, S. G. (2008). Acoustic covariants of length contrast in Japanese stops. *Journal of the International Phonetic Association,* 38(02), 167–186.

Kawahara, S. (2006). A faithfulness ranking projected from a perceptibility scale: The case of voicing in Japanese. *Language, 82*(3), 536–574.

Kawahara, S. (2012). *The phonetics of obstruent geminates, sokuon. The Handbook of Japanese Language and Linguistics: Phonetics and Phonology*.

Klatt, D. H. (1975). Voice Onset Time, Frication, and Aspiration in Word-Initial Consonant Clusters. *Journal of Speech and Hearing Research,* 18(4), 686–706.

Kraehenmann, A., & Lahiri, A. (2008). Duration differences in the articulation and acoustics of Swiss German word-initial geminate and singleton stops. *The Journal of the Acoustical Society of America,* 123(6), 4446–4455.

Lahiri, A., & Hankamer, J. (1988). The timing of geminate consonants. Jornal of Phonetics, 16(3), 327–338.

McKay, G. (1980). Medial stop gemination in Rembarrnga: a spectrographic study. Journal of Phonetics, 8, 343–352.

Liberman, A. M., & Mattingly, I. G. (1985). The motor theory of speech production revised. Cognition, 21, 1–36.

Lindblom, B. (1987). Absolute Constancy and Adaptative Variability: two Themes in *the Quest for Phonetics Invariance. In Proceedings of the 11th International Congress of Phonetic Sciences* (Vol. 3, pp. 1–18). Tallinn.

Löfqvist, A. (2005). Lip kinematics in long and short stop and fricative consonants. *The Journal of the Acoustical Society of America,* 117(2), 858–878.

Löfqvist, A. (2006). Interarticulator programming: effects of closure duration on lip and tongue coordination in Japanese. *The Journal of the Acoustical Society of America,* 120 (5 Pt 1), 2872–2883.

Löfqvist, A. (2007). Tongue movement kinematics in long and short Japanese consonants. *The Journal of the Acoustical Society of America,* 122(1), 512–518.

Löfqvist, A. (2009). Vowel-to-vowel coarticulation in Japanese: The effect of consonant duration. *The Journal of the Acoustical Society of America,* 125(2), 636–639

Ofuka, E. (2003). Sokuon /tt/-no chikaku: akusento gata-to sokuon/hisokuongo-no onkyooteki tokuchoo niyoru chigai [perception of a Japanese geminate stop /tt/: the effect of pitch type and acoustic characteristics of preceding/following vowels]. *Onsei Kenkyuu [Journal of the Phonetic Society of Japan*], 7(1), 70–76.

Port, R., Dalby, J., & O'Dell, M. (1987). Evidence for mora timing in Japanese. *Journal of the Acoustical Society of America,* 81, 1574–1585.

Ridouane, R. (2003). *Suites de consonnes en berbère : phonétique et phonologie* (Thèse de Doctorat). Université Sorbonne-Nouvelle, Paris.

Ridouane, R. (2007). Gemination in Tashlhiyt Berber: an acoustic and articulatory study. *Journal of the International Phonetic Association,* 37(02), 119–142.

Saltzman, E. L., Rubin, P. E., Goldstein, L., & Browman, C. P. (1987). Task-dynamic modeling of interarticulator coordination. *The Journal of the Acoustical Society of America, 82*(S1), S15–S15.

Sock, R., Hirsch, F., Laprie, Y., Perrier, P., Vaxelaire, B., Brock, G., Sturm, J., … (2011). An X-ray database, tools and procedures for the study of speech production. *In Proceedings of the 9th International Seminar on Speech Production (ISSP2011)* (pp. 41–48). Montréal.

Stevens, K. N. (1972). The quantal nature of speech: evidence from articulatory acoustic data. In P. B. Denes & E. E. Davis Jr (Eds.), *In Human Communication: a Unified View* (pp. 51–66). New York: McGraw-Hill Book Company.

Takeyasu, H. (2012). Sokuon-no chikaku-ni taisuru senkoo onsetsu shiin boin-no jizokujikanno eikyoo [effects of the consonant and vowel durations in the preceding syllable on the perception of geminate stops in japanese]. *Onin Kenkyuu [Phonological Studies],* 15, 67–78.

Vaxelaire, B. (1995). Single vs. double (abutted) consonants across speech rate. X-ray and acoustic data for French. In *Proceedings of the 13th International Conference on Phonetic Sciences* (Vol. 1, pp. 384–387). Stockholm.

Vaxelaire, B., & Sock, R. (1997). Laryngeal movements and speech rate, an X-ray investigation. In *Eurospeech'97 & 5th European Conference on Speech Communication and Technology*, (Vol. 2, pp. 1039–1042). Rhodes, Grèce.

TICAM'16, p. 249 - 255

Constitution des corpus pour l'analyse de discours. Transcription synchronisée à l'aide du logiciel Winpitch

Ramdane Boukherrouf **Noura Tigziri**
Laboratoire d'aménagement et de l'enseignement de la langue amazighe
Université Mouloud Mammeri de Tizi-Ouzou.
{ramdaneboukherrouf, Nora.tigziri}@gmail.com

Résumé

Notre travail consiste à présenter une application informatique faite à l'aide du logiciel Winpitch qui consiste à mettre en place une base de données numérique des corpus à exploiter dans le domaine de l'analyse de discours. Notre travail propose une application qui prend en charge la transcription synchronisée des corpus oraux (texte / son). Cette dernière contient six niveaux de représentation : le fichier son, le discours et le silence, la transcription phonétique, la notation usuelle, la traduction juxtalinéaire et la traduction sémantique.

1. Introduction

Notre contribution consiste à présenter une partie des applications informatiques de notre projet[1] (Challah, 2004) (Boukherrouf, 2006), (Boukherrouf, 2013), (Tigziri, 2014), (Tigziri, 2016) qui consiste à la mise en place d'une base de données de corpus oraux, numérisés, transcrits et annotés pour la langue amazighe qui soit destiné à différentes exploitations scientifiques par les chercheurs berbérisants. Nous avons un certain nombre de corpus qu'on exploite pour différentes applications, notamment en géographie linguistique et en dialectologie. Dans notre cas, nous nous contentons de présenter une application de constitution des corpus qui seront exploités dans le domaine de l'analyse de discours.

Notre travail s'inscrit dans le domaine de l'analyse textuelle des discours. Nous nous baserons essentiellement sur l'approche d'Adam (2011). Nous avons choisi d'inscrire notre recherche dans cette approche, parce qu'il s'agit d'une théorie qui prend en charge la production contextuelle du sens. C'est pourquoi, elle exige de travailler sur des corpus recueillis dans des situations concrètes.

1 Intitulé : Transcription synchronisée des corpus oraux, dirigé par le professeure Nora Tigziri, domicilié au Laboratoire d'Aménagement et d'Enseignement de la Langue Amazighe.

Par ailleurs, les approches textuelles considèrent la ponctuation comme partie indissociable du verbal, et ce dans la construction sémantique et la segmentation des différentes unités textuelles ; phrases, périodes, séquences, etc. Au niveau de l'oral, la prosodie joue un rôle primordial dans la structuration syntaxique des énoncés.

L'élaboration de notre application est faite à l'aide du logiciel Winpitch (Martin, 2006) qui prend en charge dans l'alignement la transcription synchronisée du corpus texte / son avec la prise en charge des différents contextes de pauses et les différents niveaux de transcriptions.

Pour mener à bien notre travail, nous l'avons structuré en cinq points afin de traiter de façon progressive la problématique.

Le premier point consiste à présenter le corpus et son contexte de production. Dans le deuxième point, nous présentons le logiciel Winpitch. Le troisième point concerne la présentation des différents contextes de pauses qui figurent dans notre corpus. Le quatrième point présente les différents niveaux de traductions et de transcriptions. Quant au cinquième et dernier point, nous le consacrons à la présentation de l'application de synchronisation texte / son.

2. Le corpus : présentation et contexte de production

Nous allons travailler sur un texte concret. Il s'agit d'un discours politique émanant d'un acteur politique kabyle. Notre texte a été enregistré dans une situation réelle : un discours du docteur Saïd Sadi, d'une durée de quarante-cinq minutes, prononcé lors du meeting tenu à l'occasion des élections législatives de 2002, le 02 mai 2002 au stade Oukil Ramdane de Tizi-Ouzou.

3. Présentation de l'outil d'analyse : Le logiciel Winpitch

Cette application sera réalisée par l'intermédiaire d'un autre logiciel : *le winpitch*[2] (figure 1) qui prend en charge les domaines de la prosodie et plusieurs niveaux d'alignement.

Figure 1. Application WinPitch

2 Un logiciel de la prosodie conçu par le professeur Philippe Martin http://alsic.org

Winpitch comporte un certain nombre d'applications dont celles qui nous intéressent pour le moment : l'analyse acoustique, l'alignement son/texte. En lançant, Winpitch, nous voyons apparaître la fenêtre ci-dessous (figure 2). En cliquant sur la touche Alignement, une fenêtre s'active à gauche :

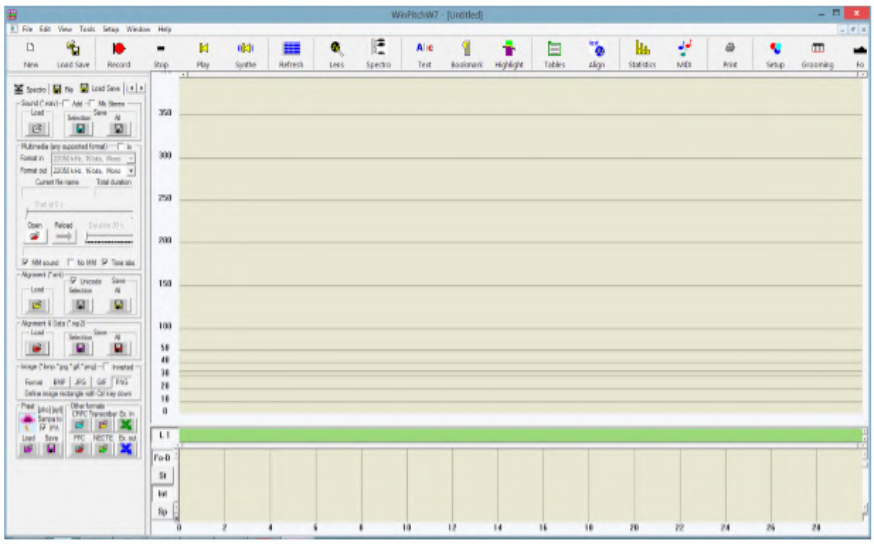


Figure 2. Fenêtre d'accueil WinPitch

Le texte : de la version sonore au texte

Comme le soulignent les philologues[3], avant d'entamer l'analyse d'un texte, il est impératif d'intervenir sur l'ensemble des versions disponibles, afin de s'assurer de son authenticité. La génétique du texte doit être introduite comme donnée préalable par la linguistique avant de procéder à l'analyse de tout texte. En d'autres termes, le chercheur ne doit pas prendre son texte comme donnée mais comme objet à construire. Cependant, pour ce qui nous concerne nous n'avons pu nous procurer qu'une version du corpus. Il s'agit d'une version sonore, enregistrée sur cassette audio par notre collègue B.S., étudiant à l'époque, présent lors du meeting. Le texte est pris tel quel, du début à la fin, il n'a subi aucun remaniement ni normalisation.

Dans les pages suivantes, nous procédons à la constitution de notre texte par sa transcription et sa traduction en langue française, et ce, pour rendre possible son analyse. Ces deux tâches font partie de notre thèse : d'une part, une réflexion théorique et méthodologique sur l'établissement du texte pour l'analyse, d'autre part, une traduction du travail (réflexion sur la traduction).

3 Nous faisons référence, notamment au travail de Rastier (2001).

Transcription et traduction du discours : un texte pour l'analyse

La transcription du corpus, comporte cinq niveaux. Le premier niveau consiste en la présentation des différents silences représentés par « S » et les différentes unités sonores du discours représentées par « D ». Le deuxième et le troisième niveaux consistent en la transcription phonétique et la notation usuelle du corpus dans la langue de l'orateur, le quatrième est réservé à la traduction juxtalinéaire. Le cinquième et dernier niveau représente la traduction libre.

Comme indiqué précédemment et pour rester le plus fidèle possible au texte oral, nous avons donc opté pour une transcription phonétique et une notation usuelle. Pour reproduire fidèlement les réalisations orales du corpus, nous avons pris en considération les faux départs, les hésitations, les pauses, représentées par (/) et les applaudissements, représentés par (A). Le choix des deux transcriptions est justifié par la prise en charge des réalisations (phonétiques) orales du corpus, et le décodage du corpus par les chercheurs linguistes non berbérophones pour la transcription phonétique et la fluidité de son décodage par les praticiens pour la notation usuelle.

Dans la transcription phonétique, nous avons reproduit globalement les graphèmes de l'Alphabet Phonétique International (API).

- Le graphème barré indique l'emphase [ṛ, ṣ, ṭ]
- Les affriquées sont transcrites avec deux graphèmes associés par une ligature [tŝ, dẑ]
- Les labiovélaires sont accompagnées d'un graphème [w] en exposant [k^w]
- La tension est transcrite avec deux graphèmes identiques [mm, ss, ll]
- A côté des sons emphatiques, les voyelles [a, i, u] sont réalisées respectivement [α, e, o].

Concernant la notation usuelle, nous avons adopté une notation à tendance phonétique pour mettre en valeur les marques de l'oral utilisées par l'orateur dans son discours.

La traduction libre est précédée par une traduction française juxtalinéaire pour indiquer la position des unités dans le texte d'origine, les conventions et les abréviations utilisées dans la traduction.

Le verbe à l'infinitif suivi de la marque aspectuelle :

- + P : Prétérit
- + A : Aoriste
- + AI : Aoriste Intensif
- + PN : Prétérit Négatif

Et des modalités préverbales et/ou postverbales :

- NR : Modalité du non Réel (modalité aspectuelle du verbe)
- AC : Modalité d'Actuel Concomitant (Modalité aspectuelle du verbe).
- ICI : Modalité d'orientation spatiale vers le locuteur.
- LA-BAS : Modalité d'orientation spatiale (éloignement)
- NEG : Négation

Le nom suivi de la marque d'état :

- EL : Etat Libre
- EA : Etat d'Annexion

Quant à la traduction libre nous avons opté pour une traduction plus ou moins fidèle au texte afin de mettre en valeur les référents culturels de la langue source dans la version traduite.

Transcription synchronisée : texte/ son

Après avoir présenté précédemment les différents niveaux de transcriptions adoptés dans l'établissement de notre corpus, le logiciel de synchronisation et les niveaux de transcription adoptées dans la présentation des corpus, ce point est réservé à la présentation de l'application de la synchronisation texte / son avec le logiciel Winpitch. En effet, la synchronisation présente six niveaux principaux :

- L1 : fichier son
- L2 : segmentation discours-silence
- L3 : transcription phonétique segmentée
- L4 : notation usuelle
- L5 : traduction juxtalinéaire
- L6 : traduction sémantique

Ci-dessous, un exemple de figure de winpitch qui présente un exemple d'une transcription synchronisé son /texte (figure 3).

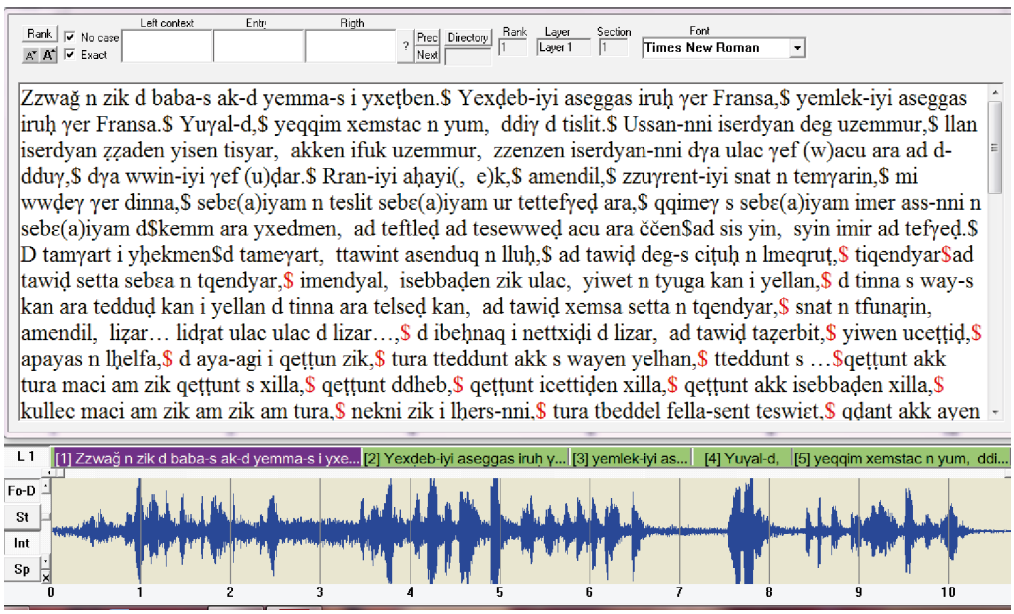
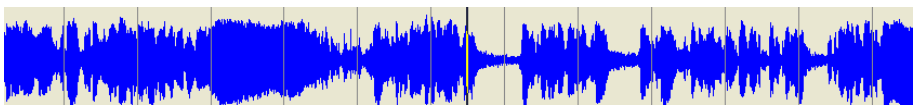


Figure. 3 Exemple d'une transcription synchronisé son /texte

Le « dollar » ($) représente les frontières de segmentation qui sont établies par l'auteur au fur et à mesure qu'il avance dans Winpitch. Cette segmentation est faite sur la base des différents silences qui figurent dans le discours de l'orateur. Les autres niveaux d'alignements sont créés successivement au-dessous du fichier son.

Ci-dessous (figure 4), les résultats de la synchronisation texte-son représentant les différents niveaux d'alignement de notre application.

- L1 : fichier son
- L2 : segmentation discours-silence
- L3 : transcription phonétique segmentée
- L4 : notation usuelle
- L5 : traduction juxtalinéaire
- L6 : traduction sémantique



D	S	D	S	D
qˆaənasiməzwuranˆɔə // isərsuħədd̪aðθafd̪isθirəfðitˆsmˆis // /[ðajənigəlˆanðiθiʁinnuvəmbəvuxəmsin //				
Qaren-as imezwura-nneɣ//isersuḥeddadtafḍstirfed-tt mmi-s///d ayen I yellan di tiɣri n novembre 45				
dire+AI-Ilsanciens+EA-nos // il-déposer+P le forgeron+EA le marteau+ELil+prendre+P-le+A fils-son /// C'est ce qui il-être+Participe dans l'appel+EL de nouvembre cinquante-quatre //				
Nos anciens disent // (quand) le forgeron dépose son marteau son fils le reprend /// C'est ce qui est contenu dans l'appel du 1er novembre cinquante-quatre //				

Figure 4 Synchronisation texte-son représentant les différents niveaux d'alignement

7. Conclusion

En guise de conclusion, notre application nous permet de prendre en considération les différents niveaux de l'oral (son et pause) et de l'écrit (transcription et traduction) qui sont des niveaux indispensables dans l'analyse textuelle des discours. En effet, notre application nous permet de mettre à la disposition des chercheurs qui travaillent dans le domaine de l'analyse de discours des textes établis avec les différents niveaux recommandés par le domaine en question.

En terme de perspective, il est nécessaire d'intégrer d'autres paramètres de l'oral, notammentles hésitations, l'accent, l'intonations qui sont indispensables pour le traitement de la structuration et la hiérarchisation du discours.

Références

Adam J-M.(2011). *Linguistique textuelle*, Paris, Armand Colin (3ème édition).

Boukherrouf R. (2006). *Structures intonatives des énoncés verbaux complexes en berbère (kabyle)* coordination et subordination, Mémoire de Magistère, Université Mouloud Mammeri de Tizi-Ouzou.

Boukherrouf R. (2013). «Le rôle de la sémantique et de la prosodie dans la distinction entre la subordination et la coordination sans marque monématique de jonction», *Faits de syntaxe amazighe*, Actes du colloque internantional organisé par le Centre d'Aménagement Linguistique, IRCAM, Rabat Maroc, pp.55-65.

Chaker S (1991). «Eléments de prosodie berbère : quelques données exploratoires », in EDB N°8, PP.5-25.

Chaker S. (2009) : «Structuration prosodique et structuration (typo-) graphique en berbère : exemples kabyles.», *in Etudes de phonétique et linguistique berbère, Hommage à Naima Louali* (1961-2005), PEETERS, Paris Louvain, PP.227-242.

Challah S. (2004). Le rôle de l'intonation entre en syntaxe. Etude de cas portant sur l'opposition de l'état (*analyse intonosyntaxique de quelques types d'énoncés*), Mémoire de Magistère, Université Mouloud Mammeri de Tizi-Ouzou.

Duez D. (1991). *La pause de la parole de l'homme politique*, Paris, Edition du CNRS, collection Son et Parole.

Grevisse M. (1980). *Le bon usage- grammaire française avec des remarques sur la langue française d 'Ajourd 'hui*, Duculot.

Martin P. (2006). WinPitch LTL, un logiciel multim´edia d'enseignement de la prosodie. Apprentissage des Langues et Syst`emes d'Information et de Communication, 2006, 08 (1), pp.95-108. <http://alsic.org>. <edutice-00109613>.

Mettouchi, A. (2003). « Focalisation contrastive et structure de l'information *en kabyle (berbère)», Mémoires de la Société de Linguistique de Paris*, PP. 179-196.

Tigziri N. (2016). Analyse textuelle à l'aide du concordancier ANTONC d'une œuvre de BélaidAt Ali (Bu yedmimen et tassa), Colloque international *«BELAID AIT ALI (1909-1950)*. Un auteur et une œuvre à (re)lire», les 24-25 avril 2016, Tizi-Ouzou (à paraître)

Tigziri Noura (2016). Analyse de l'album ISEFRA Lounis Ait Menguellet à l'aide de ANTCONC in colloque international *« Litt'Orale », De la littérature orale comme patrimoine, Regards croisés*, les 7 et 8 avril 2016, Oujda (à paraitre)

Tigzri N. (2014). La réalisation de grands corpus berbères normalisés et interopérables : enjeu culturel et enjeu d'ingénierie linguistique in Revue *ASINAG N 9*, Rabat, Maroc.

Reconnaissance des caractères Tifinaghe imprimés par une description structurelle

Youssef Ouadid Brahim Minaoui Mohamed Fakir
Department of Computer Science, Faculty of Science and Technology Laboratory of Information Processing and Decision and Support
Beni-Mellal, USMS, Morocco
yo.ouadid@gmail.com, bra_min@yahoo.fr,fakfad@yahoo.fr

Résumé

Dans cet article, un système de reconnaissance des caractères tifinagh imprimés est présenté. Il est divisé en trois étapes principales: prétraitement, extraction de caractéristiques et classification. Dans la phase de prétraitement, la qualité de l'image est améliorée en utilisant les fonctions de binarisation, normalisation et squelletisation. Dans la phase d'extraction de caractérisitques, le squelette du caractère est divisé en plusieurs segments et representé par un graphe dont les nœuds representent les points singuliers et les arrêtes representent les segments du caractère. En phase de classificaion, l'algorithme d'appariement spectral est utilisé pour comparer les graphes. Ceux-ci representent les caractères générés à partir des images de la base de données IRCAM. Des résultats experimentaux sont atteints en utilisant 3267 caractères aléatoires pour tester l'efficacité. Le système montre de bons résultats en terme de précision.

1. Introduction

La langue amazighe est parlée aujourd'hui par environ 14 millions de personnes, principalement dans le Maghreb. L'écriture amazighe est normalement écrite de gauche à droite et verticalement de haut en bas, c'est une écriture non cursive ce qui simplifie la segmentation des caractères dans une image de texte amazighe. La figure 1, illustre les caractères tifinagh adoptés par l'IRCAM.

Le système de reconnaissance optique des caractères est un outil qui facilite l'interaction homme-machine. C'est la procédure qui convertit une image de texte en un document texte modifiable par l'ordinateur ou un matériel similaire. Il est utilisé dans plusieurs domaines où le travail est basé sur les documents texte, principalement dans le bureau à des fins d'indexation ou d'archivage automatique des documents par exemple. De nombreuses recherches ont été concentrées dans la création d'un système de reconnaissance optique des caractères chinois (Mansi et Jethava, 2013) et arabe (LLorigo et Govindaraju, 2006), en particulier les caractères latins. En revanche, la reconnaissance des caractères Tifinagh, reste moins explorée.

Figure 1. Caractères Tifinaghe

Les récentes approches proposées pour la reconnaissance des caractères Tifinagh impriméssont basées généralement sur l'hybridation de plusieurs descripteurs (Oujaoura *et al.,* 2013) (El Ayachi *et al.*, 2012), classifieurs (Bencharef *et al.*, 2011) (El Kessab *et al.,* 2011), ou les deux (Oujaoura *et al.,* 2014) ; ce qui donne de très bonnes résultats en terme de taux de reconnaissance. En contrepartie, combiner plusieurs descripteurs/classifieurs augmente considérablement la complexité du système. Autre systèmes intérressants ont été proposé, (Fakir *et al.,* 2011) ont présenté un système qui se base sur l'utilisations des variétés riemanniennes et la classification par SVM (Machine à Vecteurs de Support). (Gounane *et al.*, 2011) ont proposé deux systèmes pour comparer deux méthodes de classification non supervisée, les cartes auto adaptatives et les k-plus proches voisins. Pour minimiser le taux d'erreur, le bigram a été utilisé (Gounane *et al.* , 2013). Dans une nouvelle approche, (Essaady *et al.,* 2012) ont proposé un système basé sur l'histogramme vertical et horizontal pour extraire le vecteur caracteristique afin de l'introduire dans les Modèles de Markov Caché. Une approche similaire a été proposé par (Amrouch *et al.* , 2012) dont ils ont utilisé une séquence de vecteur caractéristique directionnelle. (El Ayachi *et al.,* 2014) ont proposé un système de reconnaissance basé sur la représentation des caractères par un code brèle. Récemment, nous avons proposé des systèmes de reconnaissance basés sur la représentation structurelle des caractères tifinagh. Il s'agit de la représentation par graphes. Dans une première approche, (Ouadid *et al.,* 2014) nous avons utilisé la représentation par les matrices d'incidence puis une comparaison matricielle exacte. Le système était très rapide mais faible contre le changement de taille et de fonte. Dans une deuxième approche (Ouadid *et al.,* 2016), nous avons proposé un système basé sur l'extraction des points singuliers par la méthode de Harris, puis nous avons représenté ces points et leurs connectivités par un graphe. Ensuite, nous avons comparé ces graphes par une méthode d'appariemment spectral ce qui a donné des résultats satisfaisants. Cependant la méthode de Harris est légérement vulnérable au changement de taille. Cela produit un nombre de points d'inflexion plus que ce qu'on désire.

Dans ce travail, nous proposons un système de reconnaissance de caractères, basé sur une approche structurelle qui est les graphes. Pour une meilleur représentation par graphe, un algorithme d'extraction des points singuliers, précisemment les points d'inflexion a été proposé.Ce système est composé de trois phases principales : Prétraitement, Extraction des caractéristiques structurelles et classification (Figure 2).

L'organisation du reste de cet article est comme suit : dans la section 2, la phase de prétraitement du caractère est décrite. La section 3, traite la phase d'extraction des caractéristiques structurelles. La section 4, s'occupe de la phase de classification et enfin, dans la dernière section des résultats experimentaux sont discutés.

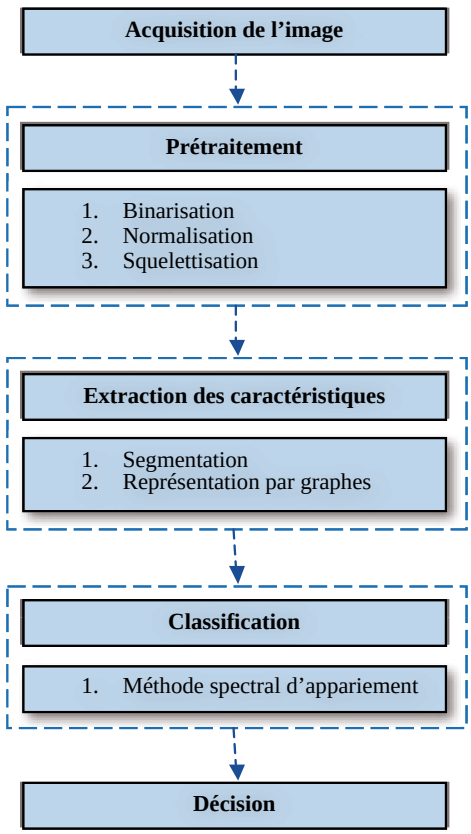


Figure 2. Système de reconnaissance de caractère Tifinagh

2. Prétraitement

Le prétraitement vise à réduire le nombre de données en gardant que les informations pertinentes. Figure 3 illustre un exemple de processus de prétraitement d'un caractère Tifinagh.

Nous avons procédé par la binarisation en utilisant la bien connue méthode Otsu (Otsu *et al.*, 1979). Le caractère est ensuite normalisé en éliminant le vide entre ces bordures et celles de l'image. L'histogramme horizontal est scanné verticalement et l'histogramme vertical est scanné horizontalement pour détecter les coins du caractère.

Le squelette du caractère est extrait en utilisant la méthode de Zhang-Suen (Zhang *et al.*, 1984). C'est un algorithme d'amincissement parallèle en deux itérations pour éliminer les pixels qui n'appartiennent pas au squelette du caractère. Cependant, dans certains cas, l'algorithme produit des squelettes d'épaisseur de deux pixels dans certains endroits. Ainsi, dans le cas du caractère numéro 25, le cercle remplit est éliminé entièrement. Pour remédier à ces problèmes nous avons proposé les solutions suivantes :

1. Appliquer la squelettisation sur les composants connexes au lieu de l'appliquer sur l'image entière. Cela nous permet de traiter le cas des composants qui sont éliminés après l'amincissement.
2. Lisser le squelette en éliminant les pixels qui n'affectent pas la connectivité et la topologie du caractère. Il s'agit de traiter les pixels dont le nombre de voisinage et le nombre des transactions 01 est différent.

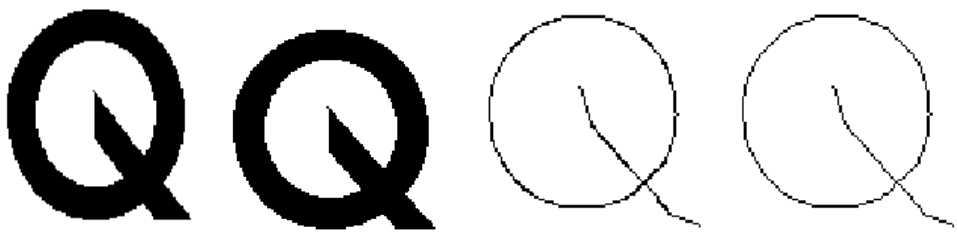


Figure 3. Exemple du processus de prétraitement des caractère Tifinagh, (a) Image Original monochrome, (b) Normalisation, (c) Squelettisation, (d) Lissage du squelette

3. Extraction des caractérisitques Structurelles

Le but de cette étape est de fournir une description synthétique du caractère. En utilisant l'approche structurelle, qui est la représentation graphique. Nous avons proposé un algorithme pour diviser le squelette du caractère en plusieurs lignes.

Extraction des points singuliers

Afin de diviser correctement le squelette du caractère en plusieurs lignes, nous avons commencé par extraire les points clés primaires qui sont les points d'intersection et les points d'extrémités. Il s'agit des pixels dont le nombre de transaction de zero (0) à un (1) dans les voisinages est égale à 1 (point extrémité) ou supérieur à 2 (point d'intersection). Figure 4.(b) illustre un exemple d'extraction des points extrémités et d'intersection. Dans le cas

des caractères circulaire, comme le caractère 1, on n'aura pas de points d'intersection ni d'extrémité. Pour remédier à ce problème, un point est ajouté comme point singulier (celui du coin haut à gauche).

L'étape suivante consiste à trouver des points d'inflexion. Tout d'abord, Nous avonsextrait les segments entre les points clés primaires (voir figure 4.(c)), cela va nous permettre d'analyser chaque segment séparément et décider si un point d'inflexion est nécessaire ou non. Nous avons proposé un critère qui permet de classer les segments en deux types de forme géométrique simple : ligne et arc. Il s'agit de calculer la longueur du segment D par rapport à la longueur de la ligne d qui relie les deux extrémités du segment. Dans notre cas, le seuil qui détermine le type de la forme du segment est 0.2.

$$\begin{cases} \dfrac{D - d}{D} \geq 0.2 & \text{le segment est un arc} \\ \dfrac{D - d}{D} < 0.2 & \text{le segment est une ligne} \end{cases} \quad (1)$$

Lorsqu'un arc est détecté, la distance orthogonale entre les éléments du segment et la ligne droite est calculée. L'élément qui possède la distance perpendiculaire maximale est désigné comme un point d'inflexion. Chaque fois un point est ajouté dans un segment, deux nouveaux segments remplacent l'ancien segment dans la liste des segments. L'algorithme utilisé pour l'extraction des points est le suivant :

1. Extraire les points d'intersections et points d'extrémité.
2. Extraire les segments reliant ces points.
3. Pour chaque segment, vérifier si c'est un arc ou ligne.
4. Pour chaque arc designer le pixel qui a la plus grande distance orthogonale comme un point singulier.
5. Mettre à jour la liste des segments.
6. Tant qu'il y a un segment de type arc répéter l'étape 3 à 5.

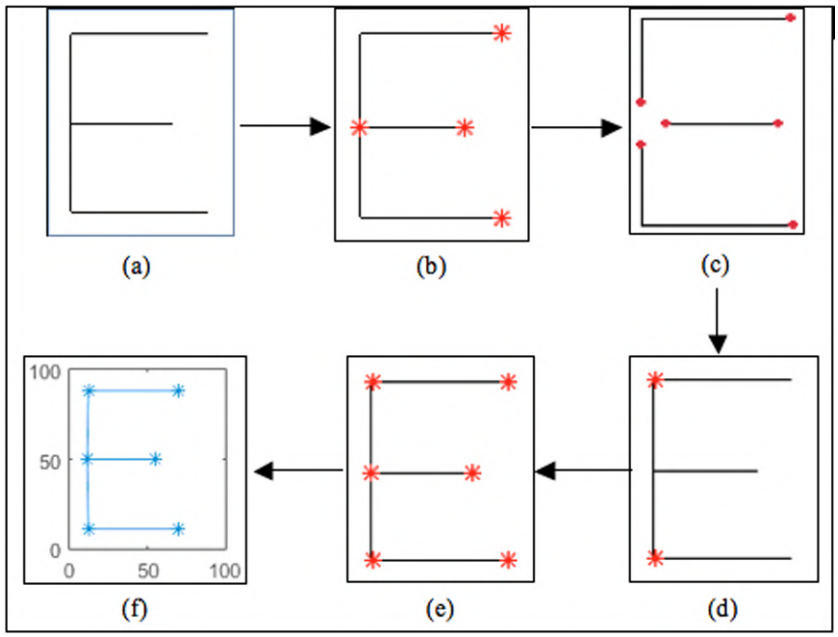


Figure 4. Exemple du processus d'extraction des caractéristiques structurelles d'un caractère Tifinagh, (a) squelette du caractère, (b) points d'intersection et d'extremité, (c) segmentation du caractère basé sur les points primaires, (d) points d'inflexion, (e) points singuliers, (f) representation du caractère par graphe

Représentation graphique

De nombreux travaux sur la reconnaissance de formes par la représentation graphique ont été réalisés (Vento, 2015). Ils sont tous d'accord sur le fait que la conversion d'une image en un graphe est une étape très importante. Le graphe doit représenter les caractéristiques de la lettre dans une large mesure.

Un graphe est une représentation mathématique formelle d'un ensemble d'objets et de leurs relations. Chaque objet est appelé un sommet. Les relations entre les objets sont appelées arrêtes. Plus formellement, nous définissons un graphe G comme une paire ordonnée de G= (V, E) où V est un ensemble de sommets, E est un ensemble d'arrêtes qui définissent la connectivité entre une paire de sommets.

La matrice d'adjacence est une façon de représenter les graphes. Il s'agit d'une matrice carrée binaire AM où le nombre de sommets | V | est sa taille. L'entrée dans la ligne i et la colonne j est non nul si et seulement si l'arrête (i, j) est dans le graphe ce qui signifie :

$$\begin{cases} AM(i,j) = 1, & \textit{sommet i et j sont connectés par une arrête} \\ AM(i,j) = 0, & \textit{sommet i et j ne sont pas connectés} \end{cases} \quad (2)$$

Dans notre cas, nous avons travaillé sur les graphes non orientés dont la direction des arrêtes n'est pas importante. Cette représentation est rapide et compacte. Cependant, il existe une redondance d'information lorsqu'on travaille avec les graphes non orientés.

Après la segmentation du caractère, les caractéristiques extraites seront représentées par un graphe non orienté, où les nœuds représentent les points singuliers et les arrêtes représentent les segments reliant ces points par paire. L'algorithme de construction graphique est procédé comme suit :

1. Soit *AM* une matrice d'adjacence de taille $n \times n$. Dont n est le nombre des points singuliers.
2. Pour chaque segment, 4 informations sont extraites : premier point singulier, second points, orientation *O* (calculé à l'aide des fonctions existante dans MATLAB) du segment, longueur *L* du segment.
3. En se basant sur ces informations, la matrice d'adjacence est construite tell que :

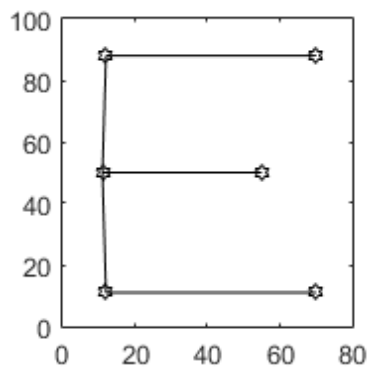
$$\begin{cases} AM(i, j) = \omega, & \textit{point clé i est connecté au point clé j} \\ AM(i, j) = 0, & \textit{else} \end{cases} \quad (3)$$

Tel que, $$\omega = 2\times L + O \quad (4)$$

Figure 5 illustre un exemple de représentation matricielle et graphique du caractère Yaḍ.

0	0	0	45	0	0
0	0	0	0	59	0
0	0	0	0	0	59
45	0	0	0	85	84
0	59	0	85	0	0
0	0	59	84	0	0

(a)



(b)

Figure 5. Représentation structurelle du caractère Yaḍ: (a) Matrice d'adjacence, (b): Représentation graphique

4. Classification

Appariemment des graphes

La méthode spectrale utilisée (Leordeanu et Hebert, 2005) vise à trouver un accord cohérent entre deux ensembles de caractéristiques P (P contient $\mathbf{n_P}$ élément) et Q (Q contient $\mathbf{n_0}$ élément). Elle est généralement utilisée pour trouver le composant principal dans un graphe. Ceci est réalisé en créant la matrice d'adjacence M, où les sommets représentent l'attribution potentielle (a = (**i**, **i'**), et $\boldsymbol{i \in P \;\; i' \in Q}$) entre deux points singuliers et les arrêtes représentent le degré de ressemblance entre ces affectations. Ils sont représentés par un poid positif si deux paires d'affectations sont convenables, autrement une valeur nulle est assignée. Dans notre cas, nous avons utilisé un mappage un-à-un de correspondance (au lieu de un à plusieurs). Ce qui signifie qu'un point singulier du premier set est assigné à un seul point à partir du deuxième set. Cette représentation prend en compte la géométrie du caractère et le degré de similarité entre les segments et les points singuliers. En observant la matrice d'adjacence M, nous pouvons extraire les propriétés spectrales qui peuvent nous aider à déterminer à quel point une paire de caractéristiques est reliée au regroupement principal. Nous continuons de rejeter la correspondance avec une faible association jusqu'à ce que nous atteignions le mappage un-à-un de la correspondance. La procédure pour construire la matrice M est la suivante :

Étant donné C un ensemble de paires où a = (**i**, **i'**) $\in \boldsymbol{C}$. Il contient toutes les affectations possibles qui respectent le mappage par correspondance un-à-un (une caractéristique de P est assignée au plus à une caractéristique de Q). Pour mesurer l'accord entre les deux caractéristiques de P et Q, une liste L est considéré comme le score associé pour chaque affectation candidat $\boldsymbol{a \in C}$ et une affinité associée pour chaque paire d'affectation (a, b) où $\boldsymbol{b \in C}$. Sur la base de cette liste, nous stockons ces scores sur la matrice $\boldsymbol{M}$ comme suit :

- Les entrées diagonales de M contiennent le score d'affectation individuelle $\boldsymbol{a \in C}$.
- Les autres entrées contiennent le score de paires d'affectation (a, b) telle que $\boldsymbol{b \in C}$.

Puisque nous avons utilisé les points d'intérêt comme caractéristiques, toutes les paires d'affectation sont des candidats. Maintenant, le problème de la correspondance est réduit à trouver le cluster C qui maximise le score inter cluster.

$$S = \sum_{a,b \in C} M(a,b) = x^T Mx \qquad (5)$$

x est un vecteur binaire tel que :

$$\begin{cases} x(a) = 1 & \text{if } a \in C \\ x(a) = 0 & \text{if } a \notin C \end{cases} \qquad (6)$$

Pour trouver le meilleur cluster $\mathbf{C}^*$, nous devons trouver le vecteur binaire $\boldsymbol{x}^*$ qui maximise le score $\boldsymbol{S}$. Plus grand est le score, meilleur est la similarité entre les caractères.

$$\boldsymbol{x}^* = \text{argmax}(x^T Mx) \qquad (7)$$

Nous avons appliqué la méthode en suivant l'algorithme décrit comme suit :

1. Construire la matrice symétrique et positive M.
2. Initialiser L par toutes les affectations possibles et x comme un vecteur nul.
3. Trouver a* = $\text{argmax}_{a \in L}$ (x^* (a)), tel que x^* est le vecteur propre principal de M.
4. Si x^* (a^*) = 0 retourner la solution x. Si non, soit x(a^*) =1 et retirer a^* de L.
5. Supprimer toutes les affectations qui sont en conflit avec a^* (correspondance un-à-un).
6. si L est vide retourner la solution x. Si non, retourner à l'étape 3 et 4.

5. Résultats et discussion

Des résultats sont obtenus à partir des expériences sur la base IRCAM composé de 3300 images différentes en style d'écriture et de taille. Dans la phase de prétraitement, la méthode d'histogramme pour la normalisation ainsi l'algorithme de Zhang-Suen pour la squelettisation sont utilisés. La figure 6, illustre le temps moyen écoulé pour le prétraitement de chaque caractère tifinagh. Le niveau de bruit dans la base de donnée IRCAM est en niveau bas/ moyen, du coup on n'a pas senti le besoin d'utiliser une méthode de réduction de bruit car la méthode d'Otsu est largement suffisante pour ce niveau de bruit.

Dans la phase d'extraction des caractéristiques structurelles, chaque caractère est représenté en se basant sur les points singuliers extraits. Notre objectif est de représenter un caractère par le minimum possible des points singuliers. Le tableau 1 illustre le nombre de points désiré pour chaque caractère. En utilisant l'algorithme proposé, nous avons réussi à avoir des résultats satisfaisants sans sacrifier le temps d'exécution. La figure 7& 8, illustrent le nombre de points singuliers ,extraits pour chaque caractère, comparé avec le nombre de points désiré,et le temps écoulé pour les extraires.

Dans la phase de classification, nous avons adopté une méthode d'appariement spectral des graphes. Il s'agit de calculer les propriétés spectrales de la matrice d'adjacence qui représente la similarité entre deux graphes. Pour mettre en valeur notre système, nous l'avons comparé avec d'autres systèmes réalisés dans notre laboratoire. Le tableau 2 illustre la précision et le temps d'exécution obtenus et comparés avec les résultats obtenus par d'autres systèmes. Concernant le taux d'erreur, un peu près 22 sur 3267 images de caractères sont mal classifiées. La plupart des cas sont relatifs au caractère 01 et 22, mais lorsqu'on analyse la base de donnée IRCAM, on remarque que c'est difficile pour l'œil humaine de différencier entre le caractère 01 en grande taille avec le caractère 22 en petite taille.

Ya	Yab	Yag	Yagw	Yad	Yaḍ	Yey	Yef	Yak	Yakw	Yah	Yaḥ	Yaɛ̣	Yax	Yaq	Yi	Yaj
4	5	7	11	3	6	10	8	5	9	5	4	5	7	5	6	6
Yal	Yam	Yan	Yu	Yar	Yaṛ	Yaǯ	Yas	Yaṣ	Yac	Yat	Yaṭ	Yaw	Yay	Yaz	Yaẓ	
6	4	2	8	4	7	5	6	7	9	5	9	4	4	8	11	

Tableau1. Nombre de points singuliers désirés pour chaque caractère

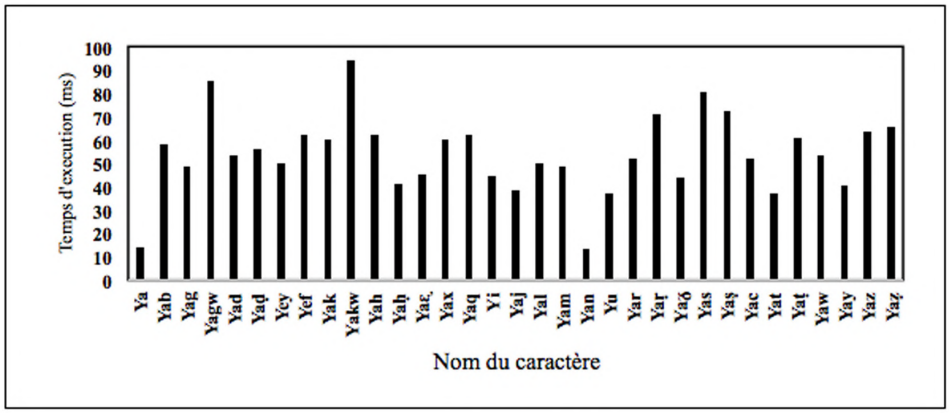


Figure 6. Temps d'execution par caractère dans la phase de prétraitement

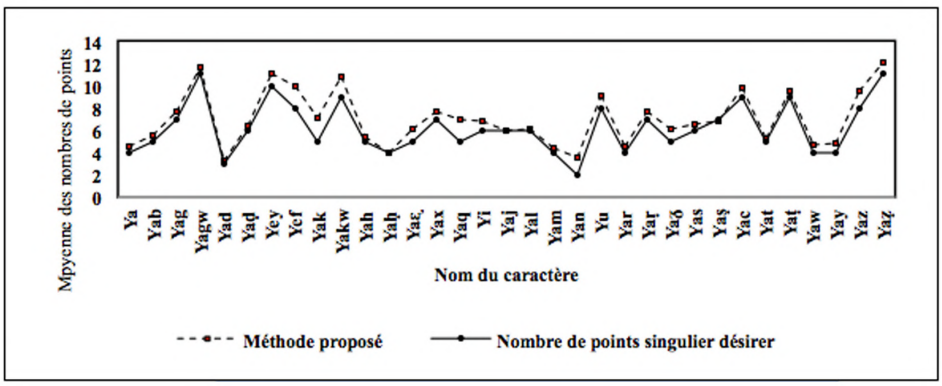


Figure 7. Comparaison des points singuliers obtenus avec ceux désirables

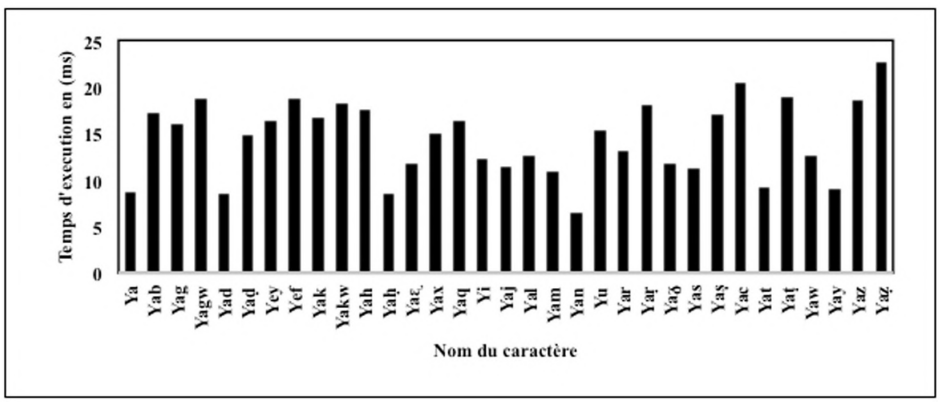


Figure 8. Temps d'execution d'extraction des caractéristiques

Système de reconnais-sance	Descripteur	Méthode structurelle proposée	Descripteur Gist	Moments de Legendre
	Classifieur	Appariement spectral des graphes	Réseaux bayésiens	Réseaux de neurone multicouche
Taux de reconnaissance (%)		99	98	81
Taux d'erreur (%)		1	2	19
Temps d'exécution (sec)		3200	5912.676	154.28 28

Tableau 2 . Comparaison des systèmes de reconnaissance des caractères tifinaghe

6. Conclusion

Dans ce travail, nous avons présenté un système rapide et précis pour la reconnaissance des caractères Tifinagh. IL est basé sur la mise en correspondance des graphes via une méthode d'appariement spectral efficace. Cette méthode utilise les propriétés spectrales de la matrice d'affinité pour calculer le score de similarité entre les graphes. Chaque graphe représente les caractéristiques structurelles du caractère qui sont les points singuliers et leurs connectivités. Les résultats des expériencesréalisés montrent que la plupart des caractères sont reconnus sans compromettre le temps d'exécution. Dans les travaux futurs, nous envisagerons appliquer le système sur l'écriture manuscrite tifinagh avec une plus grande base de données et niveau plus élevé de bruit pour rendre le système plus fiable.

Références

Amrouch M, Es-saady Y, Rachidi A and El Yassa M and Mammass D. (2012). Handwritten Amazighe Character Recognition System Based on Continuous HMMs and Directional Features. [International Journal of Modern Engineering Research (IJMER)]. Vol 2, Issue 2. Pages 436-441.

Bencharef O, Fakir M, Minaoui B and Bouikhalene B. (2011). Tifinagh Character Recognition Using Geodesic Distances, Decision Trees & Neural Networks. (IJACSA) [International Journal of Advanced Computer Science and Applications], Special Issue on Artificial Intelligence.

El Ayachi R, Fakir M, Bouikhalene B. (2012). Transformation de Fourier et moments invariants appliqués à la reconnaissance des caractères Tifinaghe. [e-TI - la revue électronique des technologies d'information]. Number 6.

El Ayachi R, Oujaoura M, Fakir M and Minaoui B. (2014). Code Braille et la reconnaissance d'un document écrit en Tifinagh. [The International Conference on Information and Communication Technologies for the Amazigh] (TICAM). Rabat.

El Kessab B, Daoui C, Moro K, Bouikhalene B and Fakir M. (2011). Recognition of Handwritten Tifinagh Characters Using a Multilayer Neural Networks and Hidden

Markov Model. [Global Journal of Computer Science and Technology]. Volume 11, Issue 15, Version 1.0.

Es-Saady Y., Amrouch M., Rachidi A., El Yassa M. and Mammass D. (2012). Réalisation d'un OCR pour l'écriture Amazighe imprimée. [5th International Conference on ICT for Amazigh].

Fakir M, Bencharef O, Bouikhalene B and Minaoui B. (2011). Tifinagh Character Recognition Using Riemannian Metric, SVM & Neural Networks. [International Journal of Advances in Science and Technology]. Vol 2, No 6.

Gounane S, Fakir M and Bouikhalene B. (2011). Recognition of Tifinagh Characters Using Self Organizing Map and Fuzzy K-Nearest Neighbor. [Global Journal of Computer Science and Technology]. Vol. 11, Issue 15.

Gounane S, Fakir M and Bouikhalene B. (2013). Handwritten Tifinagh Text Recognition Using Fuzzy K-NN and Bi-gram Language Model. [IJACSA Special Issue on Selected Papers from Third international symposium on Automatic Amazighe processing (SITACAM' 13)]. Beni Mellal.

LLorigo M, Govindaraju V. (2006). Offline Arabic handwriting recognition: a survey. [IEEE Transactions on Pattern Analysis and Machine Intelligence]. 28(5), 712-724.

Leordeanu M, Hebert M. (2005). A spectral technique for correspondence problems using pairwise constraints, [International Conference of Computer Vision (ICCV'05)] Vol. 1 (Vol. 2, pp. 1482-1489). IEEE.

Mansi S, Jethava G B. (2013). A Literature Review On Hand Written character Recognition. [Indian Streams Research Journal]. Vol. 3, issue -2.

Ouadid Y, Minaoui B, Fakir M, Abouelala O. (2014). New Approach of Tifinagh Character Recognition using Graph Matching, [The International Conference on Information and Communication Technologies for the Amazigh (TICAM)]. Rabat.

Ouadid Y, Minaoui B, Fakir M. (2016). Spectral Graph Matching for Printed Tifinagh Character. [13th International Conference on Computer Graphics, Imaging and Visualization (CGiV)]. Beni Mellal.

Oujaoura M, Minaoui B and Fakir M. (2013). Walsh, Texture and GIST Descriptors with Bayesian Networks for Recognition of Tifinagh Characters. [International Journal of Computer Applications]. Vol. 81, No.12.

Oujaoura M, Minaoui B, Fakir M and Bencharef O. (2014). Recognition of Isolated Printed Tifinagh Characters. International Journal of Computer Applications. Vol. 85, No 1.

Otsu N. (1979). A threshold selection method from gray-level histograms. IEEE Trans. Sys., Man., Cyber. Vol. 9 (1): 62–66

Vento M, (2015). A long trip in the charming world of graphs for Pattern Recognition, The Journal of the Pattern Recognition Society, Vol. 48, Issue 2 : 291–301.

Zhang T Y, Suen C Y. (1984). A Fast Parallel Algorithm for Thinning Digital Patterns. [Communications of the ACM]. Vol. 27, No 3.

Towards Internet Domain Names in Tifinagh

Abderrahman Ait Ali[1] **Anass Sedrati**[2]
[1] KTH Stockholm
abde@kth.se
[2] INPT Rabat
anass@kth.se

Abstract

As a global network of networks, the Internet is nowadays one of the most important tools for cultural and information exchange. With its increasing popularity, the Internet is used by more and more people from different cultural and particularly linguistic backgrounds. Language scripts such as Tifinagh are thus increasingly used. In order to integrate such language scripts, The International Corporation for Assigned Names and Numbers (ICANN) has created Internationalized Domain Names (IDN) which allow URLs to be used in different language scripts.The Amazigh language, with its script (Tifinagh), is currently an official language in Morocco. The Royal Institute of the Amazigh Culture (IRCAM) has as central objective to promote the Amazigh culture, and particularly Tifinagh both nationally, regionally and locally (IRCAM, 2016). In this sense, this paper aims at promoting the use of the language and its script in the Internet through integrating Tifinagh in URLs.

1. Introduction

The Internet is defined as a global interconnected system of computer networks. It can be seen as a network of sub-networks connected using a different wired or wireless communication technologies. The Internet started in the 1980s with ARPANET in the United States, the first network of interconnected regional academic and military networks. It became globally accessible when CERN launched the World Wide Web (WWW) in1989 (Stewart, 2000). At this stage and as the result of the origin of the Internet, most of the internet content was in English and Latin script. Thus, early computer systems were limited to the characters in the American Standard Code for Information Interchange (ASCII) which allows to write in a subset of Latin script (IANA, 2007). In particular, domain names that are used in the internet protocols were always in Latin script. However, little by little the number of internet users in the world increased to reach a larger and global extent (Figure 1). Due to this rise in the number of internet users, many different languages started to be used, particularly languages based on scripts other than the Latin one.

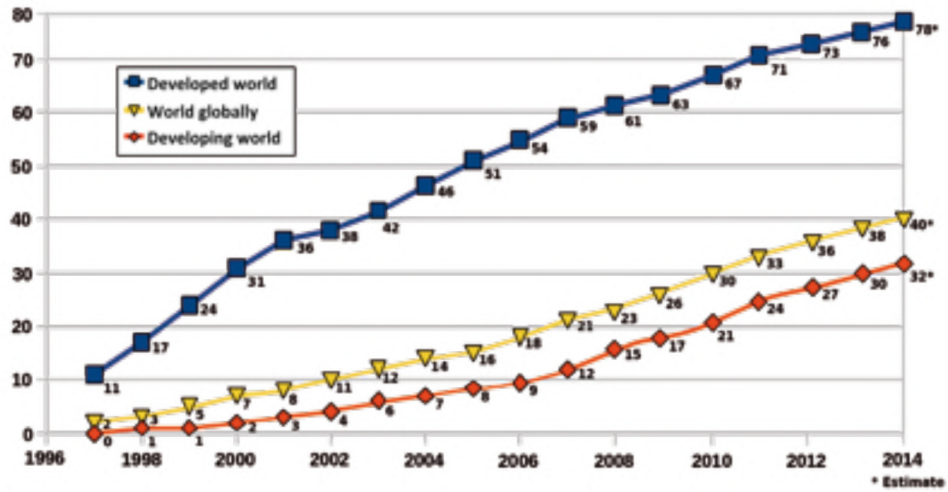


Figure 1- Internet users per 100 inhabitants (ITU, 2015)

An internalized domain name (IDN) was needed to allow internet domain names to contain language-specific scripts such as Arabic, Hebrew and Chinese. In this paper, we will discuss the case of Tifinagh, the writing script used for Amazigh languages.The case of Tifinagh is interesting and is increasingly attracting attentions especially with the recent evolutions. As of 2016, the Amazigh language is an official language in Morocco and a proposed one in the draft of the new Algerian constitution. Regardless of the common use of the Latin script and lack of knowledge of Tifinagh by native speakers, the standard Tifinagh IRCAM alphabet is used as the official script to write and teach the language in Morocco. This means that there will be a need to integrate Tifinagh alphabet in the current information systems in general and Internet domain names in particular.The integration of such a language script into the Internet domain name system faces several challenges. An administrative process, managed by ICANN, is needed in order to allow the use of Tifinagh script. Besides, some applications such as e-mail software or web browser restrict the characters which can be used as domain names. This paper will discuss the administrative process of enabling the use of Tifinagh alphabet in Internet domain names. We first present ICANN, the governing body of the Internet. Secondly, we discuss some useful technical details related to Tifinagh script. Thirdly, a description of the administrative process will be given. Finally, a set of challenges will be presented and discussed before giving some conclusions and recommendations for future work.

2. ICANN

The Internet Corporation for Assigned Names and Numbers, commonly known as ICANN, is a not-for-profit public-benefit corporation with participants from all over the world dedicated to keeping the Internet secure, stable and interoperable. It promotes competition and develops

policy on the Internet's unique identifiers. Through its coordination role of the Internet's naming system, it does have an important impact on the expansion and evolution of the Internet.

2.1. Structure

ICANN has a multi-stakeholder structural model. It treats the public sector, private sector, and technical experts as peers. The ICANN community consists of various members such as registries, registrars, internet service providers (ISPs), commercial and business interest, non-commercial and non-profit interests, governments, and individual users (ICANN, 2016).

ICANN has a very structured organization which slightly changes from time to time. As of 2016, it is managed by 16 voting members of the Board of Directors in addition to four non-voting liaisons; six representatives of different Supporting Organizations, sub-groups that deal with specific sections of the policies under ICANN's purview; an At-Large seat filled by an At-Large Organization; and the President / CEO, appointed by the Board.

Currently, ICANN has the following three Supporting Organizations:

- **GNSO**: The Generic Names Supporting Organization which deals with policy making on generic top-level domains (gTLDs).
- **ccNSO**: The Country Code Names Supporting Organization which deals with policy making on country-code top-level domains (ccTLDs).
- **ASO**: The Address Supporting Organization which deals with policy making on IP addresses.

ICANN has also different advisory committees and mechanisms to receive recommendation on the needs of other parts that are directly linked to the three Supporting Organizations. These committees include(ICANN, 2016):

- **GAC**: Governmental Advisory Committee which is composed of representatives of a large number of national governments from all over the world.
- **ALAC**: At-Large Advisory Committee which is composed of individual Internet users from around the world.
- **RSSAC**: Root Server System Advisory Committee which provides advice on the operation of the Domain Name System (DNS) root server system.
- **SSAC**: Security and Stability Advisory Committee which is composed of Internet experts who study security issues pertaining to ICANN's mandate;
- **TLG**: Technical Liaison Group which is composed of representatives of other international technical organizations that focus, at least in part, on the Internet.

The following chart (Figure 2) summarizes the organizational structure of ICANN.

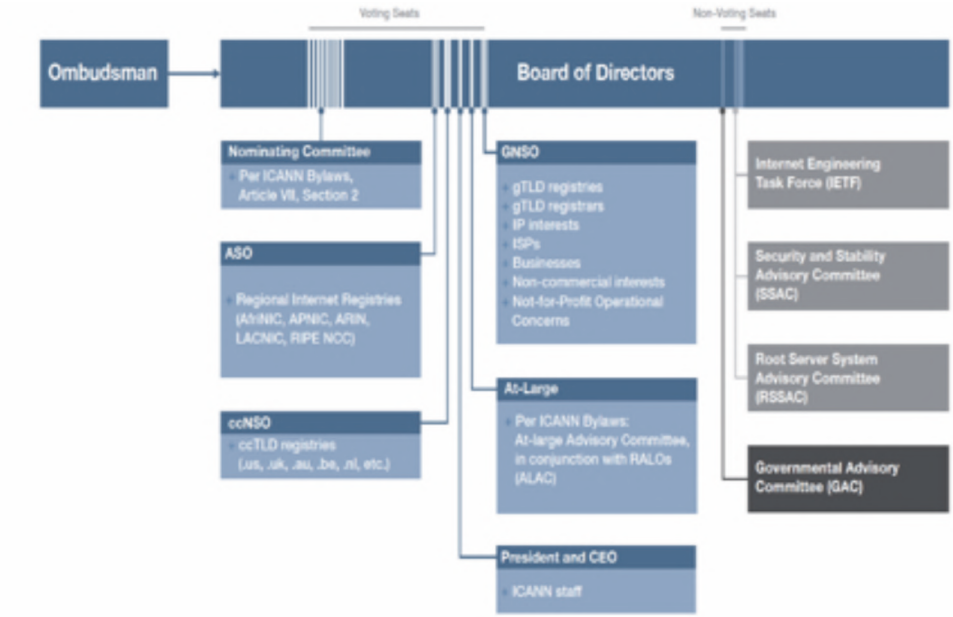


Figure 2- ICANN Organizational Structure(ICANN, 2016)

2.2. Activities

ICANN has several activities with regard to the regulation and governance of internet domains. One of its important activities is to run TLDs (Top-Level Domains) and deal with the assignment of IP addresses and ranges, ports, and other related attributes. This activity is done through an institution called IANA, Internet Assigned Numbers Authority. For instance, it is responsible for the transition from IPv4 to the new IPv6. Another important activity of ICANN is to preserve the security, stability and resiliency of the Domain Name System (DSN) and domain name registration services (ICANN, 2016). Another interesting activity of ICANN, which is important in the context of this paper, is the Internationalized Domain Names (IDNs) management. IDNs permit the global community to use a domain name in their native language or script. In this sense, ICANN focuses on the planning and implementation of the IDN Top-level Domains (TLDs) that include both ccTLDs and gTLDs. This is of paramount importance to integrate language scripts such as Tifinagh. ICANN also organizes a one week-long conference three times per year, one of them is the organization's annual General Meeting, where new board directors take their appointed seats. These meetings are held in a different location each time.

3. Tifinagh Script

The Amazigh language is the oldest attested language in the Maghreb. Its area extends east to west, from the Egyptian-Libyan border to the Canary Islands, and north to south from the southern shore of the Mediterranean to Niger, Mali and Burkina Faso.

3.1. Tifinagh Alphabet

Tifinagh is a traditional script system used to write the Amazigh language, with its various dialects. The modern Tifinagh, also known as Neo-Tifinagh is based on the traditional script of the Touareg people. It was introduced in the 20th century. A slightly modified version, called Tifinagh IRCAM is used in a number of publications in Morocco and particularly in elementary schools to teach the Amazigh language to children. Tifinagh is written from left to right.Figure 3 illustrates the Neo-Tifinagh alphabet that is used in Morocco.

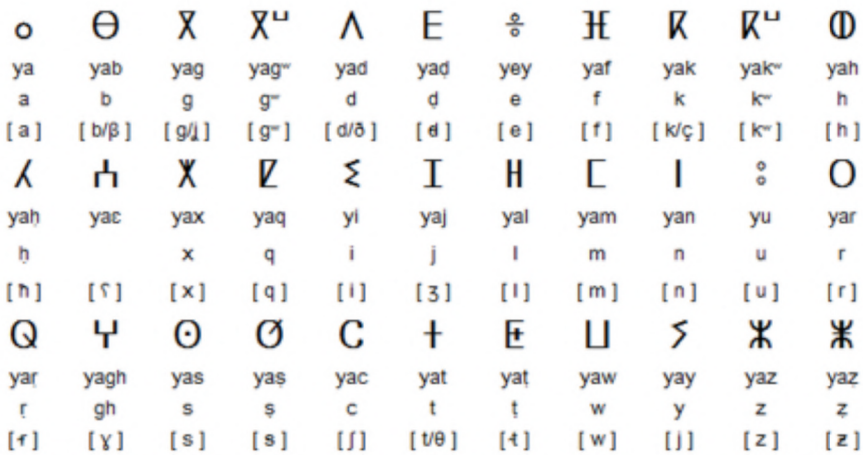


Figure 3- Neo-Tifinagh alphabet used in Morocco

3.2. Unicode

Tifinagh has currently a Unicode block in the computing industry standard for encoding and handling of texts in the world's writing systems. As of Unicode version 8.0, Tifinagh Unicode block ranges from 2D30 to 2D7F with 59 code points assigned and 21 unused reserved code points. IRCAM proposed to the International Standardization Organization (21/06/2004) the Tifinagh alphabet that was adopted. This proposal includes four subsets of Tifinagh characters:

- IRCAM basic set
- IRCAM extended set
- Other neo-Tifinagh used letters
- The modern Touareg letters whose use is attested

The following table shows the official Unicode code chart of Tifinagh block (The Unicode Consortium, 2015):

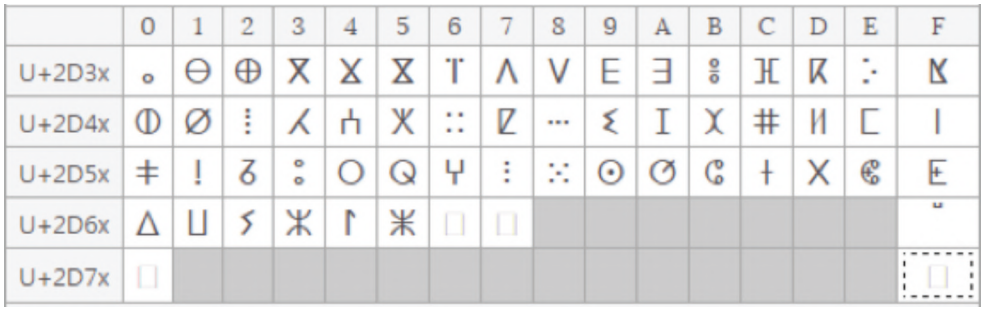


Figure 4- Tifinagh Unicode block (Unicode code version 8.0)

It is worth mentioning that Tifinagh Unicode covers the alphabets present in all the four subsets making it very flexible and easily extendable. This Unicode standardization was particularly important for the integration of Tifinagh in the new information systems such as the Internet.

3.3. Tifinagh Presence in Internet

Until recently, the use of Tifinagh in the Internet was very limited. Nowadays, it is increasingly used and supported by different platforms. For instance, Microsoft started to support Tifinagh in their Operating Systems starting from Windows 7 when Microsoft web browser Internet Explorer started to support the characters. Another example is Google browser Chrome which also supports the alphabet as of 2016. Apple in its turn started to support Tifinagh starting from its mobile operating system (iOS 9).

Figure 5- Difference between a web browser supporting (left) and not supporting (right) Tifinagh Unicode

This support trend of Tifinagh in the Internet means that more and more people will be able to read and write using the alphabet on the web. Similarly, to Arabic and languages with special writing scripts, Tifinagh needs to be supported in domain name systems so that internet users can write their web addresses in Tifinagh. This has already been done with languages as shown in the following map:

Figure 6- Some examples of internationalized domain names, adapted from(Blackbird, 2016)

Given the fact that Tifinagh is integrated in the International Unicode Standard as well as the increasing support of this Unicode by different internet services, the next step towards the integration of Tifinagh in the Internet systems will be to include it in the Internationalized Domain Names. The following sections will discuss the different steps of the administrative process and related challenges.

4. Domain Names in Tifinaghe

After having familiarized us with ICANN and seen Tifinagh technical specifications, in this part we will describe the process for the creation of Domain Names in Tifinagh script, as it is still a hanging issue.

4.1. Internationalized Domain Names

A Domain Name is a succession of characters (typically letters and numbers) separated by dots marking hierarchy. The domain name is the "address" of any page in the Internet. An example of a domain name is "event.ircam.ma". The "ma" is a Domain Name Label of the highest or top level, called Top Level Domain (TLD), "ircam" is a Domain Name Label of second level, and finally "event" is the Domain Name Label of third and lowest level. The "event.ircam.ma" can be translated into "the event that is hosted by IRCAM institute in Morocco".

As devices communicate and identify each other in a Network with IP addresses, a system was created to translate each Domain Name to its corresponding IP address. This system is called Domain Name System (DNS).

Initially, and as Internet was developed in countries adopting the Latin script, only this script in ASCII (American Standard Code for Information Interchange) format was allowed in Domain Names, and this was due to the origin of DNS which only supported Latin script in ASCII format. At that time, the rules for creating new top-level domain labels were simple:

- Labels must only be composed of letters (a-z) in the English alphabet
- Labels must have a length from 2 to 63 characters

Those rules were simple, but very restrictive. Even languages other than English using the Latin script, but not ASCII standard, were restricted. A domain name could not contain the French "é", Portuguese "ç", Swedish "å", German "ö" nor the Turkish "ğ". This issue was of course bigger in countries not adopting the Latin script at all, and created various confusing situations. In Morocco for instance, the exclusivity of Latin script forced many to write Amazigh or Arabic domain names in the Latin script, instead of Tifinagh or Arabic alphabets, as it was not possible.

As countries not adopting Latin script are many in the world, and as the aim of Internet initially is to spread and democratize knowledge to everybody, the issue of internationalizing domain names was rising more and more. It was finally solved in 2010, when a technical solution was implemented to enable the introduction of domain names in multiple scripts and languages at the top level of the DNS without destabilizing the Internet (ICANN, 2015).

After the introduction of the Internationalized Domain Names (IDNs), Top Level Domains (TLDs) were divided into 4 categories:

- **Country code Top Level Domains (ccTLDs)**: This category represents the official names of countries and territories. It includes domains in Latin script following the ASCII standard, such as .ma (Morocco) or .cat (Catalonia). The full list, set by ISO 3166, can be found here: https://www.iso.org/obp/ui/#search
- **Internationalized Domain Names of Country code Top Level Domains (IDNs ccTLDs)**: This category is similar to ccTLDs, but gathers names of countries and territories in scripts other than Latin. For instance .المغرب in Arabic or .中国 in Chinese.
- **Generic Top Level Domains (gTLDs)**: Regards any other Top Level Domain that is not a country or territory. Example: .com .org .ngo.
- **Internationalized Domain Names of generic Top Level Domains (IDNs gTLDs)**: Gathers gTLDs that are in other scripts, such as .삼성 (.Samsung) or .□□□□□(.Organization).

4.2. Current Status

In Morocco, according to the 2011 constitution, the two official languages are Tamazight and Arabic (Moroccan Government, 2011). A solution for domain names has already been implemented for Arabic by ICANN. ICANN has indeed since 2010 (and until mid-2015) approved 47 IDNs ccTLDs that cover 15 scripts and 24 languages(ICANN, 2015). Deployment is still ongoing and will continue until all scripts that ask for TLDs are covered.

The same solution ought now to be implemented for the Amazigh language and the Tifinagh script. To our knowledge and as of late 2016, the current status is that no one has already asked (or even started the process) for an Internationalized Domain Name in Tifinagh.

Our aim through this paper is to initiate the process so that in the very near future one can type the URL "ⵔⵎⵔⵎ.ⵔⵜⵜ.ⵎⵔ"[1] [2] as an alternative to "event.ircam.ma", and that it works. We hereby express our interest in bringing together experts from IRCAM with those from ICANNto make IDN in Tifinagh a reality.

In order to achieve this, the following steps need to be achieved in collaboration with ICANN:

• Create a Tifinagh Script Generation Panel

A Script Generation Panel is a panel supported by ICANN that is responsible for developing proposals that determine script-specific Label Generation Rules (LGR) for the root zone. The Amazigh speaking community, interested in developing Tifinagh script in TLDs, can engage through the Tifinagh Generation Panel. Registering in a panel is done by sending an email to idntlds@icann.org, and specifying the interest in the Amazigh language and Tifinagh script (APNIC, 2014).

This panel should contain experts in DNS, Unicode, IDNs, linguistics (Amazigh language), and domain name operations and policy. All challenges and milestones are discussed in the panel, and receive support from ICANN experts if needed. Issues can be various, and are related to the script, either regarding the technological aspect, or specific dialects, or confusion between characters, etc.

In the case of Tifinagh, and as Morocco is not the only Tamazight-speaking country in the world, the Tifinagh Script Generation Panel might also include participants from Algeria, Tunisia, Libya, Egypt, Mauritania, Canary Islands, Mali, Niger and Burkina Faso. Moreover, any person interested in Tifinagh script and Amazigh language is welcome to join the panel and participate with their time and knowledge in the project.

• Establish the Tifinagh Code Point Table

A Code Point Table is a subset of Unicode code points that is designed by members of the Script Generation Panel, and covers all characters that represent the new script that should be in the TLDs. The table gathering all code point tables is called the Maximal Starting Repertoire (MSR). ICANN has already released a first version of MSR called MSR-1. This repertoire covers nowadays 22 scripts (among which Arabic, Cyrillic, Greek, Thai…). Once the Tifinagh Panel establishes its Code Point Table and agrees on a definitive version, it can suggest the Label Generation Rules for Tifinagh. The proposal will then be raised to an ICANN experts Panel called the Integration Panel. Once accepted, the Integration Panel will integrate the LGR for Tifinagh Script into the main LGR (APNIC, 2014).

1 ⵔⵜⵜ is the abbreviation if IRCAM in Tamazight, and stands for ⵔⵉⵣⵔⵡⵉⵍ.

2 ⵜⴰⵏⴰⵎⵓⵔⵉⵜ. ⵎⵔ is our suggested ccTLD for Morocco (in Tamazight) that will be discussed further in the next section.

5. Challenges

In this section, we will cover some of the technical challenges that have to be solved, before, during and after sending the Tifinagh LGR proposal.

• Linguistic issues

As in any language, Tamazight or the Amazigh language contains a variety of dialects and differences, not only in Morocco, but also in the whole North Africa. If IRCAM has done an appreciated work of standardization of Tamazight in Morocco, this work has still to be done at a larger extend within the other Tamazight-speaking countries. The Tamhasheq dialect (spoken by Touareg) has for example characters that are specific to it in Tifinagh, and are not used in Morocco. The Tifinagh Script Generation Panel must be very cautious regarding this issue. Moreover, differences in dialects create a challenge. All confusions between dialects must be cleared in order to create a working Unicode Code Points Table. One important aspect is that all characters in the table must be part of an active and contemporary variant of the Tamazight language.

Tifinagh is the script gathering all Tamazight speakers, and enabling it should give the opportunity to all Tamazight speaking communities to use it in their TLDs. This will need a tight cooperation between communities, and even sometimes compromises in order to reach a satisfying consensus for all parts.

• Choice of names for TLDs

When applying for Top Level Domain names in Tifinagh script, the Generation Panel should at least have standards for the most commonly used TLDs. We suggest in the table below an example of ccTLDs in Tifinagh for some Tamazight speaking countries and territories.

		Script (ASCII)	in Tifinagh
Morocco	ⵍⵎⴰⵖⵔⴱ	.ma	.ⵎⵖⵔ
Algeria	ⴰⵣⴰⵢⴻⵔ	.dz	.ⴰⵣ
Tunisia	ⵜⴻⵏⵙ	.tn	.ⵜⵏ
Libya	ⵍⵉⴱⵢⴰ	.ly	.ⵍⵉ
Egypt	ⵎⵉⵙⵔ	.eg	.ⵎⵙⵔ
Mauritania	ⵎⴻⵔⵉⵜⴰⵏⵢⴰ	.mr	.ⵎⵔⵜ
Canary Islands	ⵜⵉⵔⵣⵔⵓⵉⵜⵉⵕⴰⵏⴰⵔⵢⵉⵍ	.ic (proposed, not accepted yet)	. ⵜⵜ
	ⵎⴰⵍⵉ	.ml	.ⵎⵍ
Niger	ⵏⵉⵉⵔ	.ne	.ⵏⵔ
Burkina Faso	ⴱⴻⵓⵔⴽⵉⴰⴼⴰⵙⴻ	.bf	.ⴱⵓⵔ

This suggestion is obviously subject to modifications in the future, and is given just in order to show potential ccTLDs in Tifinagh. Other suggestions can be considered such as .ⵍⵓⵔ (from ⵍⵎⵓⵔⵓⴽ) for Morocco, or .ⵣⴰⵍ for Canary Islands.

• Register a whole Domain Name in Tifinaghe

As we have seen in the previous section, a domain name is a succession of characters separated by dots. Until now, we have only seen how to implement Tifinagh script in Top Level Domains (TLDs). However, a Domain Name is more than a TLD. It contains in fact second, third (etc.) level labels that should also be in the same script. How is this done?

If we suppose that .ⵍⵖⵓ will be the official Tifinagh ccTLD for Morocco, then registering a domain name under it has to follow the same rules as when registering domain names in .ma and .المغرب domains, that are both official domain names managed by the National Agency for Regulation of Telecommunication (ANRT). In order to register the domain dame ⵜⵜⴰⵡⵉⵏ.ⵍⵖⵓ, for example, it has to be available. This can be done through checking in www.registre.ma/whois . If the domain name is available, then a registrar should be contacted in order to apply for this domain name. This implies that Moroccan registrars must also be aware of the introduction of the Tifinagh Script, and make sure that they can support hosting domain names in Tifinagh.

• Technical challenges

Unfortunately, one of the challenges still facing Tifinagh Script is technical. Not all devices and Operative systems support yet this script, but there is a significant amelioration in the last years. Difficulties observed in reading content in Tifinagh are still present in many interfaces, but are also improved gradually by the different updates each manufacturer does.As our paper is discussing domain names in Tifinagh, our technical concern is not only about output (the webpage), but also about the input means. In order to write a domain name, a keyboard or an input system is needed. Very few keyboards sold in Morocco have a native Tifinagh script, and most of users interested in writing in Tifinagh have either to use virtual keyboards or install different programs and software tools. We hope that in the future more keyboards in Tifinagh will be released (in combination with Arabic/Latin letters), because it is the input system that often motivates more people to write in a given script. This will encourage the creation of more domain names in Tifinagh, which is hopefully our common aim.

6. Conclusions

Internet is imposing itself more and more as one of the most important and powerful tools of communication in our modern era. As a universal network, it should be accessible to all human beings regardless of languages they speak. In this matter, ICANN has been deploying efforts since 2010 in order to make domain names available in international scripts, and not only in the Latin script that was initially the only access tool to the Internet. Nowadays, people can write URLs in Arabic, Cyrillic, Chinese and many other scripts. The Amazigh language deserves clearly its place in the list. Our aim through this paper was to present guidelines and steps to follow in order to allow domain names available also in Tifinagh script. In order

to achieve this symbolic milestone, it is important for us that a team of linguistic experts works together with us in order to overcome the different challenges that might face the project. Experience from other languages and support from ICANN are provided helping tools, completing them by moving forward to panel work is the next step towards having internationalized domain names in Tifinagh.

Références

APNIC. (2014). *Speak up for your language*. Retrieved from APNIC - Asia Pacific Network Information Centre: http://blog.apnic.net/2014/09/30/speak-up-for-your-language

Blackbird. (2016). *Internationalized domain names (IDN) and DNS / Bind9 problem*. Retrieved from Brainstorming of a sysadmin: http://blackbird.si/internationalized-domain-names-idn-and-dns-bind-9

IANA. (2007). *Character Sets*. Retrieved from Internet Assigned Numbers Authority - IANA: http://www.iana.org/assignments/character-sets/character-sets.xhtml

ICANN. (2015). *Linguistic Diversity in the Internet Root: The Case of the Arabic Script and Jawi.* Retrieved from ICANN - The Internet Corporation for Assigned Names and Numbers : https://www.icann.org/news/blog/linguistic-diversity-in-the-internet-root-the-case-of-the-arabic-script-and-jawi

ICANN. (2016). *Resources - ICANN. Retrieved* from The Internet Corporation for Assigned Names and Numbers - ICANN: https://www.icann.org/resources

IRCAM. (2016). *Présentation*. Retrieved from The Royal Institute of the Amazigh Culture - IRCAM: http://www.ircam.ma/?q=fr/node/620

ITU. (2015). *World Telecommunication/ICT Indicators database 2015*. International Telecommunication Union - ITU.

Moroccan Government. (2011). *Constitution*. Retrieved from Maroc.ma: www.maroc.ma/ar/system/files/documents_page/BO_5964Bis_Ar.pdf

Stewart, W. (2000). *Living Internet*. Creative Commons.

The Unicode Consortium. (2015). *The Unicode Standard, Version 8.0*. Retrieved from Unicode: http://www.unicode.org/charts/PDF/U2D30.pdf

A Dataset of Amazigh Printed Words Images

Nabil Aharrane[1] **Abdellatif Dahmouni**[2] **Karim El Moutaouakil**[3]
Khalid Satori[4]
[1] University Sidi Mohammed Ben AbedAllah
aharranenabil@gmail.com
[2] University Sidi Mohammed Ben AbedAllah
abdellatifdahmouni@gmail.com
[3] National school of applied sciences Al-Hoceima
yassirkarimimane@gmail.com
[4] University Sidi Mohamed Ben AbedAllah
khalidsatori@gmail.com

Abstract

In the absence of a public database for a wide-scale benchmarking of Amazigh Optical Character Recognition (OCR) systems, this paper aims to provide a new Amazigh Printed Word Images database (APWID). This database contains 1795 different Amazigh words rendered with an automated procedure using different Amazigh fonts, sizes and styles to generate 1148800 word images with their ground truth xml files. The database can be used as a source of training and testing sets to evaluate Amazigh OCR systems and also in other applications such as text classification systems and Amazigh characters segmentation.

1. Introduction and motivations

In recent years, the OCR remains one of the most popular research subjects due to its diverse applications such as indexing archives, documents analysis, robotics, address classification system, processing of bank check, etc. Therefore, much work has been achieved for many languages, an excellent and recent survey can be found in (Peng *et al.*, 2013).

Recently, due to the introduction of new technologies in communication and after the official introduction of the Amazigh language's teaching in the Moroccan educational system in 2003, researchers have begun to give attention to the Amazigh language trying to provide OCR systems able to achieve better performance. Rachidi et al present an overview of some released works (Rachidi *et al.*, 2014). Unfortunately, there is no common public database that makes significant the comparison of the elaborated OCR systems.

This work has as objective to present a new Amazigh Printed Word Images database (APWID). This database will serve in our research and it is publicly available (see the link in section 3) for the scientific community to enable them to evaluate and compare their OCR systems. The variability present in the data of the APWID database allow a large-scale benchmarking of multi-fonts, multi-sizes and multi-styles recognition systems of Amazigh words

The Amazigh language has existed since the earliest antiquity. It has an original writing system, Tifinagh, used and preserved to this day. In recent decades, all Amazigh groups have reclaimed this ancestral writing. Currently, the Amazigh language is spoken by about 30 million speakers in North Africa (from the oasis of Siwa in Egypt to Morocco passing through Libya, Tunisia, Algeria, Niger, Mali, Burkina Faso and Mauritania). In Morocco, where nearly 50% of people are amazigh, the Amazigh language is divided into three regional varieties with Tarifit in the North, Tamazight in Central Morocco and South-East and Tachelhit in South-West and the High (Ameur *et al.*, 2004).

Figure 1 : Tifinagh characters adopted by the IRCAM.

The official introduction of the Amazigh language's teaching in the Moroccan educational system in 2003 involves the selection of a standard common language to teach. This task was accomplished by "Royal Institute of the Amazigh Culture" (IRCAM) created in 2001 (Sadiqi, 2011). Actually, the Tifinagh-IRCAM alphabet is based on 33 characters (IRCAM, 2003) as seen in Figure 1. In the Amazigh OCR field, the characters ⵅⷛ and ⴽⷛ do not have Unicode codes, so we obtain them by a combination of characters ' ⵅ, ⴽ' with the sign of labialization ' ⷛ' that have Unicode codes. The IRCAM institute has produced a number of Unicode-encoded Tifinagh fonts which are available for free download for all platforms. This variety in fonts allows a large variability in the rendered word images.

The rest of this paper is organized as follows: Section 2 delineates all Necessary steps (data collection, rendering procedure, images description) to build the APWID database. In section 3, we present the corresponding statistics of the database and information about storage and usefulness. Finally, we conclude the paper with Section 4.

2. The APWID database

This section describes the generation procedure of the APWID database and its specifications.

2.1. Data collection

The database contains 1795 Amazigh words created from decomposable and non-decomposable words. Decomposable words are those generated from Amazigh verbs while non-decomposable ones are formed by Amazigh proper names, days, months, animals, etc. The Amazigh words were extracted from some books such as a French-Amazigh-Arabic dictionary and a Media dictionary published by the IRCAM institute. The collected words were grouped in a text file containing one Amazigh word in each line.

2.2. Sources of variability

The images of the APWID database were generated using the 16 different Amazigh fonts proposed by the IRCAM institute (Table 1). We use all these fonts to cover different complexity of shapes of Amazigh printed characters, going from simple fonts with no or few overlaps (Tifinaghe IRCAM STANDARD) to more complex fonts rich in overlaps (Tifinaghe Tazirit UNICODE). We used also different sizes for each font: 8, 9, 10, 11, 12, 14, 16, 18, 20 and 24 points. And for each font and size combination we used different styles: Plain, Bold, Italic and Bold-Italic combination.

Amazigh Word	Fond Id	Font Name
ⵜⵉⴼⵉⵏⴰⵖ	F01	Tifinaghe-Ircam Unicode
ⵜⵉⴼⵉⵏⴰⵖ	F02	Tifinaghe-IRCAM2_unicode
ⵜⵉⴼⵉⵏⴰⵗ	F03	Tamzwart Standard UNICODE
ⵜⵉⴼⵉⵏⴰⵗ	F04	Tamalout Standard UNICODE
ⵜⵉⴼⵉⵏⴰⵖ	F05	Tifinaghe-IRCAM-taromit_unicode
ⵜⵉⴼⵉⵏⴰⵗ	F06	Teddus Standard UNICODE
ⵜⵉⴼⵉⵏⴰⵗ	F07	Tassafout Standard UNICODE
ⵜⵉⴼⵉⵏⴰⵗ	F08	Tifinaghe-Tazdayt Standard UNICODE
ⵜⵉⴼⵉⵏⴰⵖ	F09	Tifinaghe-IRCAM-Agoug_unicode
ⵜⵉⴼⵉⵏⴰⵖ	F10	Tifinaghe-IRCAM-taromit2_unicod
ⵜⵉⴼⵉⵏⴰⵗ	F11	Tamzwart Noufouss UNICODE
ⵜⵉⴼⵉⵏⴰⵗ	F12	Tamalout Noufouss UNICODE
ⵜⵉⴼⵉⵏⴰⵗ	F13	Teddus Noufouss UNICODE
ⵜⵉⴼⵉⵏⴰⵗ	F14	Tassafout Noufouss UNICODE
ⵜⵉⴼⵉⵏⴰⵗ	F15	Tifinaghe-Tazdayt Nouffouss UNICODE
ⵜⵉⴼⵉⵏⴰⵖ	F16	Tifinaghe Tazirit UNICODE

Table 1. Different Amazigh fonts proposed by IRCAM

The used fonts, sizes and styles guaranty a wide variability of the image database.

2.3. Rendering procedure

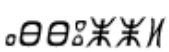
The word images were generated automatically by rendering the word text in images using a java program, so, noise and artifacts present in scanned images are not present in the image database. For each word text, we used different combinations of fonts, sizes and styles and the word image was rendered with the text anti-aliasing filtering implemented in the java standard library by RenderingHints class. The algorithm presented in the Figure 2 describes the followed rendering procedure to generate the APWID database.

```
For each word
   For each police
     For each size
       For each style
          Generate the word image file
          Generate the ground truth xml file
       End For
     End For
   End For
End For
```

Figure 2. Algorithm to generate the APWID database

2.4. Ground truth description

A detailed description for each image word of the APWID is attached by an XML file reporting ground truth information about the word sequence of characters, as well as information about the image and the rendering settings. Figure 3 illustrates an image word and its attached XML file.



```
<?xml version="1.0" encoding="UTF-8"?>
<wordimage id="IMG_4">
  <content transcription="ⵔⵙⵙⵎ" nPaws="5">
    <paw utf="\u2D30" frequency="1">ⴰ</paw>
    <paw utf="\u2D31" frequency="2">ⴱ</paw>
    <paw utf="\u2D53" frequency="1">ⵓ</paw>
    <paw utf="\u2D65" frequency="2">ⵥ</paw>
    <paw utf="\u2D4D" frequency="1">ⵎ</paw>
  </content>
  <font name="Tifinaghe-IRCAM-Agoug_unicode" fontStyle="Italic" fontSize="20" />
  <specs encoding="png" width="98" height="22" />
  <generation renderer="java" filtering="antialiasing" />
</wordimage>
```

Figure 3. Example of the ground truth XML file

This file is composed of 4 markups to provide all necessary information about the word image such as content, font, image specifications and the generation procedure:

Content: this element provides the transcription of the Amazigh word, the number of pieces of Amazigh word (nPaws), that are characters, and sub-elements for each Paw giving its correspondent utf-8 code and its appearance frequency in the word.

Font: in this element, we have information about the font used (name, size and style) to generate the word image.

Specs: this element indicates the encoding of image, its width and its height.

Generation: in this element, we give some additional information about the rendering procedure.

3. Database statistics and utilization

This section is devoted to present the APWID statistics and its corresponding information about storage and usefulness.

3.1. Statistics

The APWID database contains 1795 words composed from 10453 characters and rendered in different combinations of 16 fonts, 10 sizes and 4 styles. Table 2 reports the APWID statistics.

	Number of words	Number of characters
	1795	10453
16 fonts * 10 sizes * 4 styles		
Total	1148800	6689920

Table 2: The APWID statistics

As shown in Table 2 the APWID database is composed of 1148800 word images files, each word image is described by a ground truth xml file. The database contains in total 6689920 characters, their appearance frequencies in the database are distributed as seen in Table 3.

Character	Frequency in the APWD database	Character	Frequency in the APWD database
ⵔ	1312000	ⵉ	44160
ⵣ	140160	ⵇ	135040
ⵥ	3840	ⵀ	275840
ⵓ	336640	ⵛ	400000
ⵜ	636800	ⵍ	133120
ⵙ	128000	ⵅ	31360
ⵚ	417920	ⵄ	58240
ⵠ	529280	ⵞ	120960
ⵂ	13440	ⵏ	107520
ⵅ	33280	ⵍ	341760
ⵆ	82560	ⵣ	84480
ⵙ	477440	ⵄ	28800
ⵟ	190720	ⵒ	32640
ⵃ	160000	ⵒ	119040
ⵅ	211200	ⵕ	77440
ⵜ	7040	ⵓ	19200

Table 3: Characters appearance frequencies in the APWID database

3.2. Storage

The database takes about 900 Mo of disk space and is publicly available for free download via the following link: https://goo.gl/eCJKjF. As shown in Figure 4, the database files are organized in 16 directories representing the 16 Amazigh fonts, each font directory contains 10 other ones for the 10 font sizes and each size directory contains 4 directories for the 4 different font styles.

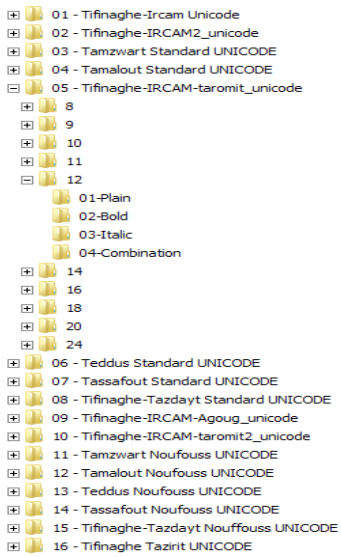


Figure 4: APWID Database structure in disk

3.3. Utilization

The APWID database can be used as training/testing sets in many applications as OCR systems, text classification systems and characters segmentation algorithms. We proposed some protocols as summarized in Table 4 to test the impact of the variability of the data in the database on the tested systems.

Protocol	**Training Set** **Train (font, size, style)**	**Testing Set** **Test (font, size, style)**
APWID1	Train (F14 ,01, P)	Test (F14 ,01, P)
APWID2	Train ((F01,F03,F05,F12,14) ,(06), (P, I))	Test ((F02,F04,F08,F12,14) ,(09), (P,I))
APWID3	Train((F01,F14,16) ,(02), (P,I))	Test ((F01,F14,16) ,(02), (B,BI))
APWID4	Train ((F-11F9,13,20) ,(15), P)	Test ((F-11F9,13,20) ,(15), P)
APWID5	Train (All, All, All)	Test (All, All, All)
APWID6	Train ((F01, F05, F08, F,(09 20,24)), All)	Test ((F01, F05, F08, F,(09 20,24)), All)

Table 4. Database testing protocols

These protocols use the notations Train (font, size, style) and Test (font, size, style) to define the training and testing conditions where:

- **Font**: the font ids as indicated in Table 1;
- **Size** : defines the sizes used in points ;
- **Style**: the style used where P, B, I and BI are for Plain, Bold, Italic and Bold & Italic.

The defined protocols have well-defined objectives and are as follow:

- **APTWID1**: This is the basic one given that there are no mismatched between the training and testing sets conditions. The performance of the OCR systems should be the highest possible;
- **APTWID2**: This one is to test the ability of systems to recognize unseen fonts.
- **APTWID3**: This protocol aims to evaluate the capability of systems to treat unseen styles.
- **APTWID4**: in this protocol, we measure the systems capability to recognize unseen sizes;

- **APTWID5**: This protocol is a global one where all data is used for experimentation;
- **APTWID6**: The last protocol is destined to the text classification systems to identify Amazigh text.

The database can be used also to test different characters segmentation algorithms in overlapped fonts such as 'Tifinaghe-IRCAM-taromit2_unicode' where the segmentation algorithm using the vertical projection histogram (Figure 5) cannot deal.

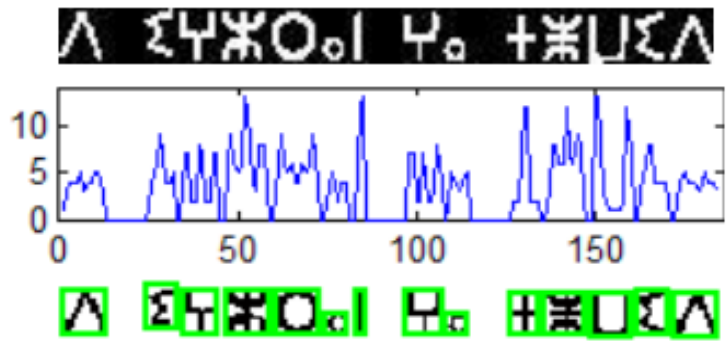


Figure 5. Characters segmentation by histogram projection (ES Saady et al., 2011)

The APWID database users are free to create their own combinations of training and testing sets according to their own needs by benefiting from the variability of the data.

4. Conclusion

In this work, we have presented a new Amazigh Printed Word Images Database consisting of 1148800 different word images and their attached ground truth XML files to provide a common database for a large-scale benchmarking of the OCR systems. The database can be used to create multiple combinations of training and testing sets while benefiting from the wide variability of the database data in term of fonts, sizes and styles. The APWID database is publicly available via Internet.

References

Ameur M., Bouhjar A., Boukhris F., Boukouss A., Boumalk A., Elmedlaoui M., Iazzi E. and Souifi H. (2004): "Initiation à la langue amazighe", Publications de l'Institut Royal de la Culture Amazighe, Manuels N.1, pp. 9.

Es Saady Y., Rachidi A., El Yassa M. and Mammass D. (2011): "Amazigh Handwritten Character Recognition based on Horizontal and Vertical Centerline of Character", International Journal of Advanced Science and Technology, Vol. 33, pp. 33-50.

Institut Royal de la Culture Amazighe. (2003) : "Proposition de codification des tifinaghes", Rabat, Morocco.

Peng X., Cao H., Setlur S., Govindarju V. and Natarajan P. (2013): "Multilingual OCR research and applications: An Overview", Proceedings of the 4th International Workshop on Multilingual OCR, ACM, New York, NY, USA. Article No.1.

Rachidi A., Eddahibi M., Essaady Y. and Amrouch M. (2014): "Amazigh Characters Automatic Recognition: Overview and Prospects", International Journal of Scientific & Engineering Research, Vol. 5, Issue 11, pp 797-803.

Sadiqi F. (2011): "The Teaching of Tifinagh (Berber) in Morocco", Handbook of Language and Ethnic Identity: The Success-Failure Continuum in Language and Ethnic Identity Efforts, Vol. 2, Oxford University Press, pp:33-44.

TICAM'16, p. 291 - 300

Toward a Handwritten Recognition System Using Canonical Representation for Multi-Script Documents

Yassine El Malki Youssef Es-Saady Driss Mammass
IRF-SIC Laboratory, IbnoZohr University
Agadir, Morocco
{y.elmalki; y.essaady ; mammass}@uiz.ac.ma

Abstract

We present in this paper a system of multilingual handwriting recognition based on canonical representation. After the word/character image preprocessing, the skeleton is transformed only into vertical and horizontal strokes, which is called the canonical representation. Then, the word/character is represented by a vector of its intersection pixels' values, those values, which are depends on the pixel's 8-surrounding neighbors. Finally, an algorithm is applied to match the character's vector with a codebook containing the vector of each character of the language. The vector that represents the character will be used in the classification step, and the system will be applied to databases of multi-script documents, which contains texts in Latin, Arabic, and/or Tifinagh.

1. Introduction

Libraries around the world have a big amount of documents. By scanning those documents their content can be accessible everywhere and by everyone. However, searching manually in scanned document, page by page seems difficult task for those who seek specific information. The solution is to use Optical Character Recognition (OCR) which is the process of transforming the image of text into text known by the computer, thus can easily be used as ASCCII code through searching information. This area of research has attracted many researches in many languages especially for Latin, Chinese and Arabic scripts (Bozinovic and Srihari 1989), (Bazzi et al 1999), (Zhang et al 2009), (Agrawal et al 2009), (Radwan et al 2013), (Es-saady et al 2014). The variability nature of the handwriting scripts made the recognition very challenging.

Many local languages such Amazigh have integrated the information systems, thus, in recent years, many researchers have been interested in Amazigh Character Recognition such as (Skoutni 2003),(Zenkouar et al 2004), (Es-saady et al 2010), (Es-saady et al 2011), (Es-saady et al 2014), because the Amazigh character needs a specific processing.

In North of Africa, the local Tifinagh writing system, which is a system of the Amazigh language, is widely used. The Tifinigh script is among the oldest script in the world, it

started from 3rd century BC. This script has known many changes from its original form as many languages like the Arabic script who changed from its original script (adding dots and diacritics marks). This script is found in the stones and tombs in some historical locations in Morocco, Algeria, Tunisia and the Tuareg areas in the Sahara.

The Amazigh alphabet which is called "Tifinagh-IRCAM", adopted by the Royal Institute of the Amazigh Culture, was officially recognized by the International Organization of Standardization (ISO) as the basic multilingual plan (Zenkouar 2014). Figure 1 represents the repertoire of Tifinagh which is recognized and used in Morocco with their correspondents in Latin characters. The number of the alphabetical phonetic entities is 33, but Unicode codes only 31 letters plus a change of a few letters to form the two phonetic units: ⵯ (g^w) and ⴽ (k^w).

Figure. 1. Tifinagh characters adopted by the Royal Institute of the Amazigh Culture with their correspondents in Latin character

The Amazigh have their own writing system since the ancient times (Gaci 2011). However, the amazigh people had used the alphabet and/or the language of the dominant people, who were in interaction with them, in the writing of the amazigh documents. Three systems have been used to transcript the Amazigh script(Skounti& al. 2003):

- The Tifinagh as an authentic alphabet in the Libyc inscription since the ancient times.
- The Arabic script, after the arrival of the Arabic in the end of 6th century.
- The Latin since the 19th century, by the colonial's scientists and later on by national researchers.

Some documents that are present in libraries have multiple languages in the same document as shown in Figure 2 extracted from a dictionary of Taoureg language(Foucauld 1951). This dictionary has both Amazigh and Latin Scripts; the Amazigh words are also written in Latin characters.

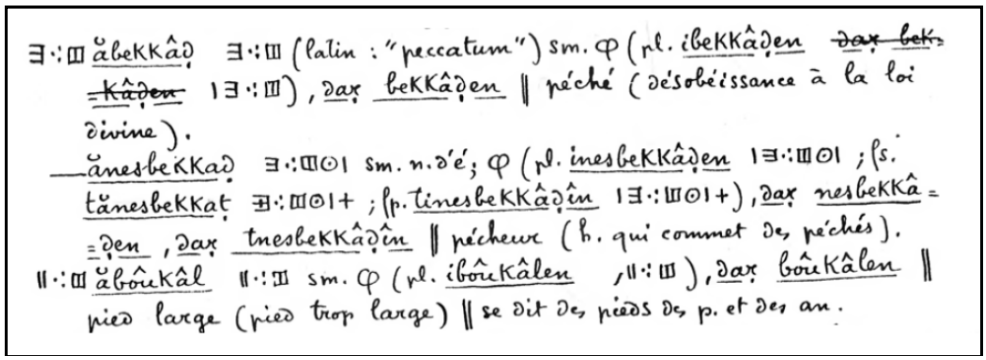


Figure 2. An Extract of Touareg dictionary (Fouclaud 1951)

Many works aimed to process multilingual documents, for example (Ambekar et al 2013) presented an OCR for printed English and Devanagari text, they used KNN for classification on 10 samples of each character of the two languages (610 samples, 260 for English and 350 for Devanagari) and they reached 95% for English text and 93% for the Devanagari Text. (G.S Lehal 2013) used a OCR for Gurmukhi and English, first they identify the script nature either Gurmukhi or English, then they used the proper OCR for each language, the test were held on 4 sets that were constructed based on 105 pages' images (76 Gurmukhi pages and 29 English Pages), they obtained different rates depending on the used set. (Tan 1998) described an automatic method for identification of Chinese, English, Greek, Russian, Malayalam and Persian text. (Das et al 2011) has used an OCR system for Telugu, English and Hindi, they extracted 8 features and then used a KNN Classifier, and they obtained 93% accuracy.

To process multilingual documents, we used a system script-free proposed by (Al Abodiand Li 2014) with some alterations, which can perform on different languages in the same text image. The scheme of the system is presented in figure 3 below.

The paper is organized as following: section 2 represent the preprocessing step, the third section describes the process of the canonical representation, finally in the section 4 we present a conclusion and some future work.

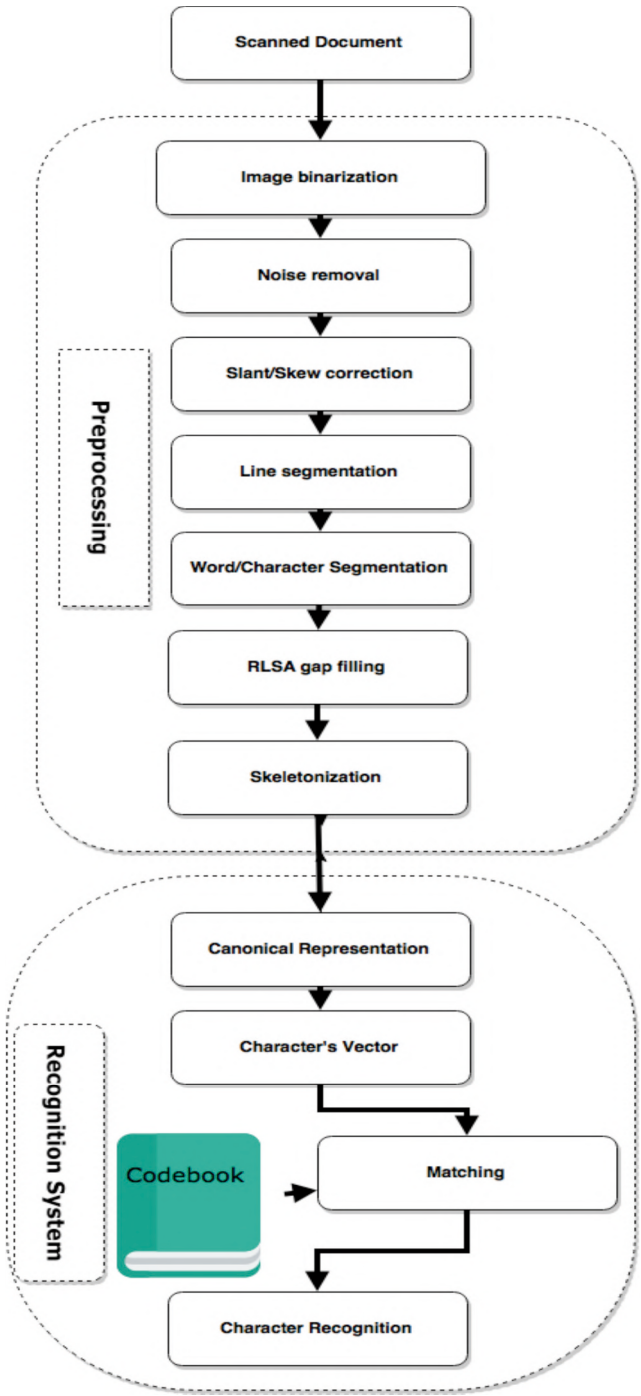


Figure 3. Proposed system scheme

2. Preprocessing

Such as was presented in Figure 3, the procedure of pre-processing which refines the scanned input image includes several steps: Binarization, noises removal and image resizing.

We used the (Otsu 1979) method for binarization. This method of thresholding is performed as a preprocessing step to remove the background noise from the picture prior to extraction of characters and recognition of text. This method performs a statistical analysis of histograms to define a function to be maximized to estimate the threshold.

The database images present gaps between pixels. This is due to the writing style or the type of the pen used. To fill the gaps, we used the widely known algorithm Run Length Smoothing Algorithm (RLSA) in both vertical and horizontal direction. For instance, if we have a sequence of values and the RLSA threshold is for example 4 all zeros between two ones becomes 1 if their number is less than the fixed threshold.

1001000010000010001 becomes 1111111100000111 1

After, a skeletonization step is applied on the image to have a pixel width of one pixel only. Figure 4 shows an example of Amazigh character and the result of skeletonization step.

a) *b)* *c)* *d)*

Figure 4. a) original image of Yab, b) binarization,c) RLSA gap filling, d) skeletonization

3. Canonical Representation

3.1 Pixel Value

We represent each pixel by a code that is obtained by summing the values of the 8 surrounding pixels, which is similar to Freeman coding. The pixel of the top has the position 0, and we move clockwise to attribute the position of the other pixels, so the pixel in the top right has the value 1, the pixel in the right has the position 2 and so one we end up with the last pixel in the top left that has the position 7. The value of the pixel is 2pixel_position. The figure 5 represents the coding scheme of the surrounding pixels.

There are 256 unique forms of surrounding pixels calculated as above. The following formulate implies that we have 256 of all geometric possible combination of the 8 neighboring pixels:

$$\sum_{i=1}^{8} C_8^i = 256$$

The figure 6 below presents some pixels with their codes:

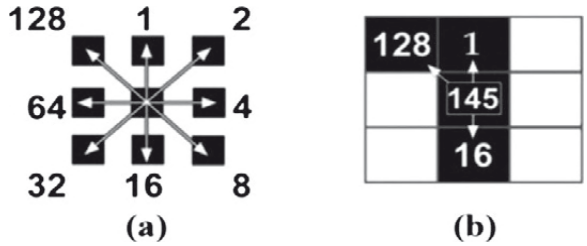


Figure 5. a) the value of the surrounding pixels b) code of the pixel is 145 = 128+1+16

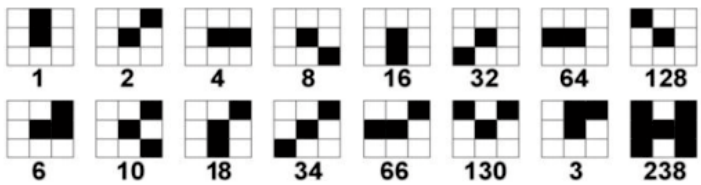


Figure 6. Some pixel codes

3.2 The Canonical Form

Scripters have deferent writing styles, this produces a problem for OCRs because the shape, the width, and the height of the characters change from one writer to another. To reduce this variety of characters forms we used the canonical representation introduced by (abode et al 14), the main concept is to have only horizontal and vertical strokes by applying processing techniques on the pixels of the image and depending on their values we got the canonical representation as described by the following algorithm:

Algorithm 1: Charcter Canonical Form

Input: An image of one-pixel width of the Amazigh character
Output: Canonical Form of the Amazigh Character
1 **Start** from the upper right corner of the image
2 **repeat**
3 **Follow** each white pixel (Foreground pixels) of the image, beginning from upper right.
4 **Using** the pixel values, we change the pixels to horizontal or vertical lines only by examining both the values of the pixel been processed and the neighboring pixels.
5 **until** *no pixel in the diagonal position*
6 **return** *Canonical Form*

For example, in Figure 7 below we begin the examination of the pixels from the top left to the bottom right of the image. In this case, the first white pixel is located in (4, 12) (column

4, line 12) with 18 as value. This value means that this pixel has an upper right pixel and a bottom pixel. The upper right pixel does not belong to the same line or the same column as the processed pixel, so we must follow the northeast direction until finding a pixel with no pixel in his upper or upper right, and have no pixel on its left pixel (17, 4). The pixels in this path will have the value of 255, and then the will be deleted. For each deleted pixel, we draw pixel in the 4 column (column of the starting pixel), and another one in the line 4 (the line of the ending pixel).

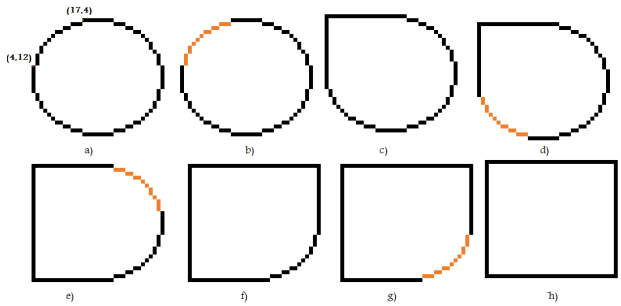


Figure 7. Example of the canonical form transformation of the Tifinagh character Ya

We continue the same process over and over until deleting all diagonal pixels from c to h. The output is the canonical form of the character. The figure 8 presents some Tifinagh characters and their canonical form and the figure 9 present an example of the canonical form Arabic word.

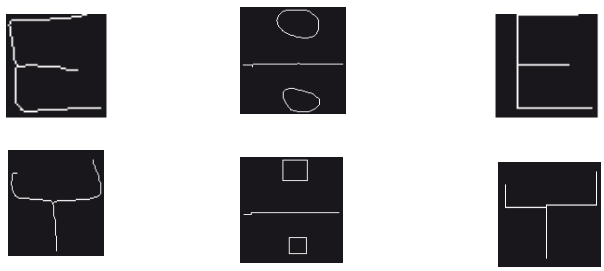


Figure 8. Example of the canonical form of some Tifinagh characters

Figure 9. Example of the canonical form Arabic word مارث *from IFN-INIT Database (Pechwitz et al 2002)*

3.3 The Character's vector

The vector value is constructed from the values of starting and the intersections pixels of the character. The starting is the upper right of the character canonical from previously explained, for example the vector of the character Yadd shown below is (64,64,64,20,21,5).

The following table presents some of the amazighe characters and their correspondent vector:

Character	Vector
ⵄ	(80,65,20,5)
ⵂ	(80,81,65,20,21,5)
ⵆ	(80,81,80,81,20,69,20,69,20,5)
ⵏ	(64,16,69,4)
ⵎ	(64,64,64,20,21,5)
ⵃ	(81,1,16,69,81,1)
ⵅ	(64,64,20,81,5,80,21,65,4)
ⵠ	(80,65,80,1,84,69,20,5,1)
ⵖ	(16,65,69,1,16,5)
ⵔ	(80,65,84,69,20,5)
ⵛ	(64,16,21,1)
ⵉ	(64,64,84,69,4,4)
ⵣ	(80,65,69,20,5)
ⵀ	(16,81,1,16,21,1)
ⵔ	(64,64,20,5)
ⵏ	(16,1)
. . .	. . .
ⵥ	(80,65,80,65,20,5,20,5)

Table1. Amazighe characters and their correspondent vector

4. Conclusion and perspectives

In this paper, we presented a system that is multi-scripts. This system can be applied to different languages present in the same image of the document. The system is tested on some of the characters of the AMCHD database (Es-Saady et al 2011), for example Yadd 91,53% ,Ya has 80,13%, Yan 96,84% recognition rate this with exact matching of the codebook vector and the character vector. However, characters with diagonal strokes present deferent vectors this because depending on stroke orientation its canonical representation will be vertical for one character and horizontal for another one. This problem can be solved using the SVM classifier with K-fold validation. In future work, we improved the method of canonical form. In addition, we will experiment our approach on a database of multi-script documents.

References

Skounti, A., Lemjidi, A., Nami, M. (2003), Tirra aux origines de l'écriture au Maroc, Publications de l'Institut Royal de la Culture Amazigh, 2003, Rabat.

Aarti, G., Ambekar, C., Hinge, S., Samidha, S., Kulkarni, S. (2013)."Bilingual OCR for Printed English and Devnagari Text". Indian Journal of Research. ISSN - 2250-1991.

De Foucauld, C. (1951)."Dictionnaire touareg-français : dialecte de l'Ahaggar". Paris :Imprimerie Nationale de France, 1951-1952, tome III (L-OU), pp. 0971-1547.

Elsayed, R. (2013). "Hybrid of Rough Neural Networks for Arabic/Farsi Handwriting Recognition". (IJARAI) International Journal of Advanced Research in Artificial Intelligence. Vol. : 2, No. 2, pp. 39-47.

Lehal, G.S (2013). "A Bilingual Gurmukhi-English OCR Based on Multiple Script Identifiers and Language Models". International Workshop on Multilingual OCR. ISBN 978-1-4503-2114-3.

Bazzi, I., Schwatz, R., Makhoul, J. (1999). "An Omni font Open-Vocabulary OCR System for English and Arabic". IEEE Transactions on pattern Analysis and Machine Intelligence. vol. 21, no 6. pp. 495-504.

Zenkouar, L. (2004). "L'écriture Amazighe Tifinaghe et Unicode", in Etudes et documents berbères. Paris (France). n° 22 pp. 175-192

Swamy, M. D., Reddy, C.R.K., SandhyaRani, D., Govardhan, A. (2011). "Script identification from Multilingual Telugu, Hindi and English Text Documents". International Journal of Wisdom Based Computing. Vol. 1 (3)

Pechwitz, M., Snoussi Maddouri, S., Margner, V., Ellouze, N., Amiri, H. (2002). "IFN/ENIT - database of handwritten Arabic words". In Proceedings of CIFED pages 129–136, 2002.

Otsu, O. (1979). "A threshold selection method from gray-level histograms". IEEE Trans. Sys, Mach., Cyber. vol. 9, pp. 62-66.

Agrawal, P., Hanmandlu, M., Brejesh, L. (2009). "Coarse Classification of Handwritten Hindi Characters". International Journal of Advanced Science and Technology. Vol: 10 pp. 43-54.

Bozinovic, R. M., Srihari, S. N. (1989). "Off-line cursive script word recognition". IEEE Trans. on Pattern Anal. Mach. Intell. vol. 11, no. 1, pp. 68-83.

Tan, T. N. (1998). "Rotation Invariant Texture Features and their use in Automatic Script Identification". IEEE Trans. On PAMI, pp. 751-756,

Es-Saady, Y. , Rachidi, A. , El Yassa, M., Mammass, D. (2011). "AMHCD: A Database for Amazigh Handwritten Character Recognition Research". International Journal of Computer Applications (0975 – 8887 IJCA). Vol: 27(4), pp.44-49, ISBN:978-93-80864-53-2

Es-Saady, Y., Amrouch M., Rachidi, A., El Yassa M., Mammass, D. (2014). "Handwritten Tifinagh Character Recognition Using Baselines Detection Features". International International Journal of Scientific & Engineering Research. Vol: 5, ISSN 2229-5518

Es-Saady, Y., Amrouch M., Rachidi, A., El Yassa M., Mammass, D. (2014). "Handwritten Tifinagh Character Recognition Using Baselines Detection Features". International Journal of Scientific & Engineering Research. Vol: 5, ISSN 2229-5518

Gaci, Z. (2011), Quel système d'écriture pour la langue berbère (le Qabyle), Mémoir de magister, 2011.

Zhang, Z., Jin, L., Ding, K., Gao, X. (2009). Character-SIFT: "A Novel Feature for Offline Handwritten Chinese Character Recognition". 10th International Conference on Document Analysis and Recognition. vol: pp. 763-767.

Al Abodi, J. , Li, X. (2014). "An effective approach to offline Arabic handwriting recognition". Comput Electr Eng. Vol:40, pp. 1883-1901.

ⵜⵙⵉⵡⵍⵜ ⵜⴰⵣⵡⴰⵢⵉⵡⵍⵜ ⵅⵜ ⵜⵛⴰⵅⵙⵢⵜ ⴰ ⵜⵙⵣⵉⵖⵓⵖⵉⵙⵎⵙⵉ ⵏ ⵙⵉⵢⵛⵙⵙⵓⵏ ⴰ ⵙⵛⵎⴰⵡⵍⵉⵏ.
ⴰⵎⴰ ⵏ ⵜⵍⵖⵓⵙⵍⵉ ⵅⵜ ⵣⵇⴰⵄ ⵏ ⵙⵓⵅⵖⵛⵏ : ⴰⵓⵉⵛⵖⴰ ⵓ ⵜⵙⵣⵉⵖⵓⵖⵉⵙⵜ, ⵙⵓⵖⵅⴰⵛ ⵙⵛⵖⵄⵄⵖⵍⵍ ⴰ ⵙⵣⵛⵖⵖⵛⵏ ⵏ ⵜⵖⵜⵉⴰⵎⵎⵜ, ⴰⵓⵛⵣⵉ ⵏ ⵍⴰⵍⴰⵉ, ⴰ ⵜⵖⵣⵅⴰ ⵜⴰⵓⵙⴰⴰⴰⵏ ⵏ ⵙⵓⵣⵣⵉⵏ.

ⵜⵙⵍⵖⵓⵙⵍⵉ ⴰⴰ ⵓⵖⵛⵍⵍ ⴰⴰⵓⴰⵍ ⵏ ⵙⵛⴰⵓⵓⴰⵍⵍ ⴰ ⵙⵛⴰⵓⵅⵖⵜⵏ ⴰ ⵙⵅⵅⵖⵎⴰⵍⵍ ⵙⵏⴰⵛⵖⵓⵏ ⴰ ⵙⵅⴰⴰⴰⵢⵎⴰⵍⵍ ⵠ ⵥⵙⵠⵓ ⵏ ⵜⵛⴰⵓⵓⴰⵏⵙⵏ ⵜⵙⵛⵖⵖⵛⵏⵙⵏ ⵜⵖⵓⵏⵙⵓⵙⵏ ⵅⵜ ⵜⵖⵜⵉⴰⵎⵥⵙⵏ ⵜⵙⵠⴰⵛⴰⵏⵙⵏ, ⵉⵉⵙⵢ ⴰⵣⵣ ⵅⵜ ⵜⵖⵜⵉⴰⵎⵥⵜ ⵜⴰⵛⴰⵅⵙⵢⵜ

* * *

الأعمال المنشورة في هذا الكتاب والتي قدمت في إطار الدورة السابعة للندوة الدولية حول الأمازيغية وتكنولوجيا المعلومات والتواصل، تتعرض لمواضيع رئيسية مثل: التعلم بوساطة التكنولوجيا ، الموارد والأدوات اللغوية، معالجة الكلام،ومقاربات التعرف الضوئي على حروف تيفيناغ. توضح هذه الأعمال مساهمات الباحثين والمهنيين الوطنيين والدوليين العاملين في مجال العلوم الرقمية المطبقة على اللغات الطبيعية، وخاصة اللغة الأمازيغية.

* * *

Les travaux retenus dans cette 7ème édition de la conférence internationale sur les Technologies d'Information et de Communication pour l'Amazighe (TICAM) et publiés dans cet ouvrage abordent quatre grands axes, à savoir : l'apprentissage médiatisé par la technologie, les ressources et outils linguistiques, le traitement de la parole et la reconnaissance optique des caractères. Ces travaux illustrent les contributions des scientifiques, chercheurs et professionnels nationaux et internationaux qui œuvrent dans le domaine des sciences du numérique appliquées aux langues naturelles et notamment la langue amazighe.

100 DH